



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM FARMACOLOGIA

Fatima Carolina Hamze

**Inteligência artificial para revisões sistemáticas em farmacologia: desenvolvimento de
um protocolo para extração de dados pré-clínicos de estudos primários**

FLORIANÓPOLIS

2026

Fatima Carolina Hamze

Inteligência artificial para revisões sistemáticas em farmacologia: desenvolvimento de um protocolo para extração de dados pré-clínicos de estudos primários

Dissertação submetida ao Programa de Pós-Graduação em Farmacologia da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Mestra em Farmacologia.

Orientadora: Profa. Dra. Cilene Lino de Oliveira

FLORIANÓPOLIS

2026

Hamze, Fatima Carolina

Inteligência artificial para revisões sistemáticas em farmacologia: desenvolvimento de um protocolo para extração de dados pré-clínicos de estudos primários / Fatima Carolina Hamze ; orientadora, Cilene Lino de Oliveira, 2026.

100 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências Biológicas, Programa de Pós-Graduação em Farmacologia, Florianópolis, 2026.

Inclui referências.

1. Farmacologia. 2. Farmacologia. 3. Revisão Sistemática. 4. Síntese de Evidência. 5. Inteligência Artificial. I. Lino de Oliveira, Cilene. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Farmacologia. III. Título.

Fatima Carolina Hamze

Inteligência artificial para revisões sistemáticas em farmacologia: desenvolvimento de um protocolo para extração de dados pré-clínicos de estudos primários

O presente trabalho em nível de Mestrado foi avaliado e aprovado, em 10 de fevereiro de 2026, pela banca examinadora composta pelos seguintes membros:

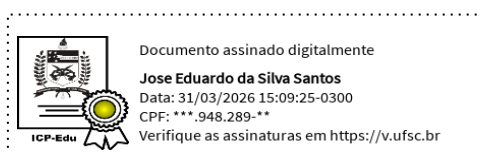
Prof. Dr. Edroaldo Lummertz da Rocha
Universidade Federal de Santa Catarina

Profa. Dra. Luz Elena Durán Carabali
Universidade Federal de Santa Catarina

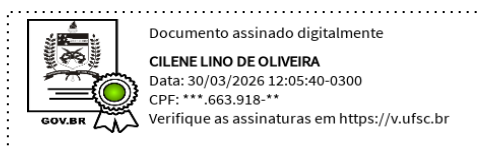
Dra. Ana Belén Ramos Hryb
Massachusetts College of Pharmacy and Health Sciences

Dra. Juliana Aparecida Bolzan
Universidade Federal de Santa Catarina

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestra em Farmacologia.



Prof. Dr. José Eduardo da Silva Santos
Coordenação do Programa de Pós-Graduação



Profa. Dra. Cilene Lino de Oliveira
Orientadora

Florianópolis, 2026

AGRADECIMENTOS

À Universidade Federal de Santa Catarina (UFSC), que calorosamente me acolheu como um segundo lar. Ao Centro de Ciências Biológicas (CCB), em especial aos departamentos de Ciências Fisiológicas e de Farmacologia, agradeço por terem sido o cenário da minha formação e pesquisa, proporcionando a estrutura e o conhecimento fundamentais para este trabalho.

Ao Programa de Pós-Graduação em Farmacologia, pela formação enriquecedora. À Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC) pelo suporte financeiro através do edital 18/2024 para bolsas de pós-graduação. Às demais agências de fomento CAPES, CNPq e Fundação Alexander von Humboldt, pelo apoio financeiro que viabiliza a existência do Laboratório de Neuroendocrinologia do Comportamento (LabNec) e de seus projetos.

À minha orientadora, Profa. Dra. Cilene Lino de Oliveira, por toda a dedicação e pelo incentivo constante a ideias inovadoras. À banca examinadora, agradeço pela disponibilidade e pelas valiosas contribuições na avaliação deste trabalho.

Aos colegas do LabNec, agradeço pela parceria. Um agradecimento especial à Juliana e à Tamires, pelo auxílio com as revisões e pela paciência em olhar meus códigos e planilhas. Aos amigos de formação externos ao laboratório, obrigada pela sabedoria compartilhada e pelos momentos de descontração.

Às minhas amigas, agradeço pelo apoio incondicional e por estarem ao meu lado, incentivando e corrigindo quando necessário. Àquelas que estão distantes, obrigada por fazerem cada reencontro valer a pena. Ao meu parceiro, agradeço pelo cuidado e pela compreensão nos momentos em que o trabalho me tomou o tempo por completo. À minha família, pelo suporte contínuo. Ao meu pai, que me acompanha em alma. E a Deus, por ser sempre um guia.

RESUMO

O uso de modelos animais na pesquisa biomédica é essencial para o avanço da farmacologia, mas enfrenta desafios críticos de reprodutibilidade e translação clínica. Diante do fenômeno da “publômica”, ou o crescimento exponencial de dados científicos que torna a síntese manual humana virtualmente impossível, as revisões sistemáticas (RS) consolidaram-se no topo da pirâmide de evidências e como aliadas dos princípios dos 3Rs (Substituição, Redução e Refinamento). No entanto, a eficiência das RS é limitada pelo gargalo metodológico da extração manual de dados, uma etapa lenta, exaustiva e sujeita a erros. Este estudo propôs e validou um protocolo computacional híbrido, utilizando as linguagens R e Python integradas a Grandes Modelos de Linguagem (LLMs), especificamente o ChatGPT e o Gemini 2.5 Pro, para automatizar a extração de variáveis farmacológicas em estudos pré-clínicos sobre a neurogênese hipocampal induzida por antidepressivos. A metodologia consistiu em um processo de engenharia de prompt iterativa para mitigar a natureza estocástica das IAs, aplicando restrições contra alucinações e padronizando a saída de dados sob a estratégia PICO. Para garantir a integridade estatística, foi aplicado um algoritmo de resolução de entidades para o alinhamento de estudos e funções de tratamento de dados para desagregar experimentos agrupados e padronizar termos e mitigar as variações terminológicas. Os resultados mostraram a superioridade temporal da automação: enquanto a extração manual de 26 artigos demandou aproximadamente 26 horas de trabalho ativo distribuídas em 12 semanas, os modelos de IA realizaram a tarefa em uma média de 10 segundos por ciclo. O Gemini 2.5 Pro destacou-se pela fluidez do processo e a capacidade processar a amostra integral em ciclo único devido à sua janela de contexto massiva, enquanto o ChatGPT foi limitado por restrições comerciais de upload diário. A validação técnica revelou que a acurácia é dependente da natureza do dado. O Gemini 2.5 Pro apresentou desempenho superior, com F_1 -scores acima de 80% na maioria das variáveis e concordância Kappa quase perfeita para espécie (100%), linhagem (98%) e tipo de fármaco (97%). Em contrapartida, ambos os modelos enfrentaram dificuldades em variáveis quantitativas e lógicas, como doses e regimes de injeção, com o ChatGPT atingindo apenas 36% de F_1 -score na identificação de frequências de tratamento. O erro predominante foi a omissão (falso negativo) em vez da invenção ou alucinação (falso positivo), muitas vezes resultando em classificação equivocada de dados existentes como não reportados (NR) pela IA. Além disso, observou-se a reprodução de vieses sistêmicos aonde, por vezes, as IAs retornaram informações que são mais frequentes na literatura ao invés da informação de interesse na amostra de artigos aplicados. Conclui-se que a IA deve ser adotada como um “auxiliar” em um modelo híbrido, onde o trabalho pesado de mineração é automatizado, mas a auditoria final permanece humana. Sob a essa ótica, o protocolo valida a tecnologia como meio de acelerar revisões tradicionais e revisões vivas, ajudando a evitar o uso redundante de animais em pesquisas cujas respostas já se encontram dispersas no dilúvio de dados científicos que já estão publicados.

Palavras-chave: *grandes modelos de linguagem, extração de dados; princípio dos 3Rs; síntese de evidências; automação.*

ABSTRACT

The use of animal models in biomedical research is essential for the advancement of pharmacology but faces critical challenges regarding reproducibility and clinical translation. Faced with the phenomenon of “publomics”, or the exponential growth of scientific data that makes manual human synthesis virtually impossible, systematic reviews (SRs) have established themselves at the apex of the evidence pyramid and as allies of the 3Rs principles (Replacement, Reduction, and Refinement). However, the efficiency of SRs is limited by the methodological bottleneck of manual data extraction, a slow, exhaustive, and error-prone stage. This study proposed and validated a hybrid computational protocol using R and Python integrated with Large Language Models (LLMs), specifically ChatGPT and Gemini 2.5 Pro, to automate the extraction of pharmacological variables in preclinical studies regarding the hippocampal neurogenesis of antidepressants. The methodology consisted of an iterative prompt engineering process to mitigate the stochastic nature of AIs, applying constraints against hallucinations and standardizing data output under the PICO strategy. To ensure statistical integrity, a best-fit algorithm was developed for study alignment, along with data-handling functions to disaggregate grouped experiments and standardize terms to mitigate terminological variations. The results demonstrated the temporal superiority of automation: while the manual extraction of 26 articles required approximately 26 hours of active work distributed over 12 weeks, the AI models performed the task in an average of 10 seconds per cycle. Gemini 2.5 Pro stood out for its process fluidity and the ability to process the entire sample in a single cycle due to its massive context window, while ChatGPT was limited by commercial daily upload restrictions. Technical validation revealed that accuracy is dependent on the nature of the data. Gemini 2.5 Pro showed superior performance, with F_1 -scores above 80% for most variables and almost perfect Kappa agreement for species (100%), strain (98%), and drug type (97%). Conversely, both models struggled with quantitative and logical variables, such as doses and injection regimens, with ChatGPT achieving only a 36% F_1 -score in identifying treatment frequencies. The predominant error was omission (false negative) rather than invention or hallucination (false positive), often resulting in the AI misclassifying existing data as “not reported” (NR). Furthermore, the reproduction of systemic biases was observed, where AIs occasionally returned information that is more frequent in the literature instead of the specific information of interest within the sampled articles. It is concluded that AI should be adopted as an “assistant” in a hybrid model, where the heavy mining work is automated, but the final audit remains human. Under this perspective, the protocol validates the technology as a means to accelerate traditional and living reviews, helping to avoid the redundant use of animals in research whose answers are already dispersed within the deluge of published scientific data.

Keywords: *large language models; data extraction; 3Rs principles; evidence synthesis; automation.*

LISTA DE FIGURAS

Figura 1. Estratégias de aprimoramento do eixo translacional em pesquisas pré-clínicas	17
Figura 2. Pirâmide de evidência científica	18
Figura 3. Etapas da condução de revisões sistemáticas e meta-análises	20
Figura 4. Hierarquia dos domínios da inteligência artificial	26
Figura 5. Representação esquemática do processo de geração de texto em um LLM	27
Figura 6. Mecanismo de resolução de entidades (entity resolution).	36
Figura 7. Estrutura de uma matriz de confusão	40
Figura 8. Comparação estrutural de estudos extraídos por artigo antes (a) e após o alinhamento (b) com o ChatGPT	60
Figura 9. Comparação estrutural de estudos extraídos por artigo antes (a) e após o alinhamento (b) com o Gemini	61
Figura 10. Análise de desempenho global por variável da extração automatizada do Gemini vs. ChatGPT	63
Figura 11. Análises de concordâncias para variáveis single-label da extração automatizada com o ChatGPT (a) e Gemini (b).	64
Figura 12. Matriz de confusão para a espécie (a), sexo (b) e linhagem (c) para os dados extraídos no GPT	65
Figura 13. Matriz de confusão para a espécie (a), sexo (b) e linhagem (c) para os dados extraídos no Gemini	66
Figura 14. Matriz de confusão para antidepressivos (a), dose de antidepressivo (b), número de injeções de antidepressivo (c) e duração do tratamento (d) para os dados extraídos no ChatGPT	68
Figura 15. Matriz de confusão para antidepressivos (a), dose de antidepressivo (b), número de injeções de antidepressivo (c) e duração do tratamento (d) para os dados extraídos no Gemini	69
Figura 16. Matriz de confusão para marcador de neurogênese (a), dose de BrdU (b), número de injeções de BrdU (c) e tempo de sobrevivência (d) para os dados extraídos no GPT	70
Figura 17. Matriz de confusão para marcador de neurogênese (a), número de injeções de BrdU (b), dose de BrdU (c) e tempo de sobrevivência (d) para os dados extraídos no Gemini	71
Figura 18. Concordância para protocolos de estresse extraídos manualmente vs. ChatGPT (a)	

e Gemini (b) 73

Figura 19. Concordância para testes comportamentais extraídos manualmente vs. ChatGPT (a)

e Gemini (b) 75

LISTA DE TABELAS

Tabela 1. Comparativo entre os Grandes Modelos de Linguagem (LLMs).....	42
Tabela 2. Tempo gasto na extração de dados da amostra de artigos.....	46
Tabela 3. Estrutura da tabulação dos dados extraídos pré e pós-tratamento	46
Tabela 4. Variáveis para a população (P) das extrações Manual, GPT e Gemini.....	47
Tabela 5. Variáveis para a intervenção (I) das extrações Manual, GPT e Gemini	50
Tabela 6. Variáveis para o desfecho (O) das extrações Manual, GPT e Gemini	54
Tabela 7. Alinhamento de dados para pareamento de estudos e identificação de omissões e alucinações	58

LISTA DE QUADROS

Quadro 1. Código desenvolvido para o tratamento de dados	35
Quadro 2. Escala de interpretação do Índice Kappa (κ)	39
Quadro 3. Estrutura inicial do prompt para extração de dados em revisão sistemática de farmacologia com base na ferramenta PICO.....	43
Quadro 4. Estrutura final do prompt para extração de dados em revisão sistemática de farmacologia com base na ferramenta PICO.....	43

LISTA DE ABREVIATURAS E SIGLAS

ACT	Teste de Actímetro
AMI	Amitriptilina
BWT	Teste de Travessia na Barra
CIT	Citalopram
CMS	Estresse Crônico Moderado
CUMS	Estresse Crônico Moderado Imprevisível
CUS	Estresse Crônico Imprevisível
CYL	Teste do Cilindro
DES	Desipramina
DL	Aprendizado Profundo
ELS	Estresse Precoce
EPM	Teste de Labirinto em Cruz Elevado
ESC	Escitalopram
FCP	Teste de Paradigma de Condicionamento do Medo
FLV	Fluvoxamina
FLX	Fluoxetina
FST	Teste de Nado Forçado
GST	Teste de Força de Preensão
IA	Inteligência Artificial
IMI	Imipramina
LLM	Grande Modelo de Linguagem
LPS	Modelo de Depressão Induzida por LPS
ML	Aprendizado de Máquina
MS	Estresse de Separação Maternal
MWM	Teste do Labirinto Aquático de Morris
NEST	Teste de Construção de Ninho
NLP	Processamento de Linguagem Natural
NSF	Teste de Alimentação Suprimida pela Novidade
OFT	Teste de Campo Aberto
PRX	Paroxetina
ROT	Teste do Rotarod ou Cilindro Giratório
RSMA	Revisões Sistemáticas e Meta-análises

SBX	Teste da Caixa de Transição
SIR	Estresse por Isolamento Social
SIT	Teste de Interação Social
SPL	Teste do Borrifo
SPT	Teste de Preferência por Sacarose
SRT	Estresse por Contenção
TST	Teste de Suspensão pela Cauda
UTT	Teste de Resposta Inata ao Som
WIRS	Estresse por Contenção e Imersão em Água

SUMÁRIO

1 INTRODUÇÃO	16
1.1 Modelos animais em pesquisa biomédica	16
1.2 Revisões sistemáticas e meta-análises como alternativa à experimentação animal	18
1.3 Automatização de revisões sistemáticas e meta-análises	21
1.4 Inteligência artificial como ferramenta auxiliar para extração de dados	24
2 HIPÓTESE	28
3 OBJETIVO GERAL	29
3.1 Objetivos específicos	29
4 METODOLOGIA	29
4.1 Computador, <i>softwares</i> e pacotes utilizados	30
4.2 Padronização e desenvolvimento do protocolo de extração	30
4.2.1 <i>Seleção do modelo LLM</i>	31
4.2.2 <i>Engenharia de prompt</i>	31
4.3 Aplicabilidade do protocolo de extração em uma amostra representativa	32
4.3.1 <i>Características da amostra</i>	32
4.3.2 <i>Extração automatizada e protocolo de execução</i>	32
4.4 Validação da acurácia e eficiência da extração automatizada	32
4.4.1 <i>Tratamento e padronização de dados</i>	33
4.4.2 <i>Alinhamento e pareamento de estudos</i>	35
4.4.3 <i>Análises de concordância e métricas de desempenho</i>	36
5 RESULTADOS	41
5.1 Estruturação e padronização do protocolo de extração automatizada	41
5.2 Extração de dados quantitativos e qualitativos GPT ou Gemini vs. Manual	45
5.2.1 <i>Extração de variáveis para a população</i>	46
5.2.2 <i>Extração de variáveis para a intervenção</i>	50
5.2.3 <i>Extração de variáveis para o desfecho</i>	54
5.2.4 <i>Comparação estrutural de estudos extraídos por artigo pré e pós-alinhamento</i>	58
5.3 Análises estatísticas e métricas de desempenho das ferramentas automatizadas	62
5.3.1 <i>Métricas de recuperação para os dados extraídos pelos modelos de IA</i>	62
5.3.2 <i>Métricas de concordância para os dados extraídos pelos modelos de IA</i>	63
6 DISCUSSÃO	75
6.1 Extração de dados automatizada para revisões sistemáticas na farmacologia	75
6.2 Desenvolvimento do protocolo para extração de dados com modelos de IA	77

6.3 Validade estrutural dos dados extraídos pelos modelos de IA	79
6.4 Validade do conteúdo dos dados extraídos pelos modelos de IA.....	80
6.5 Aplicação de modelos de IA em uma revisão sistemática de neurofarmacologia.....	83
6.6 Limitações e perspectivas futuras.....	86
7 CONCLUSÃO.....	87
REFERÊNCIAS.....	89

1 INTRODUÇÃO

1.1 Modelos animais em pesquisa biomédica

O uso de modelos animais se tornou frequente nas últimas décadas, trazendo avanços notáveis na Farmacologia e outras áreas biomédicas. O uso é justificado principalmente pela notável semelhança fisiológica e anatômica entre humanos e animais, especialmente mamíferos (Mukherjee et al., 2022). Estes modelos fornecem uma população de estudo relativamente homogênea para a investigação de condições experimentais similares as doenças e seus possíveis tratamentos (Ibrahim et al., 2022). Modelos animais são também usados para estudar a segurança e a eficácia terapêutica dos fármacos, aspectos essenciais para a aplicação em humanos (Mukherjee et al., 2022). Contudo, dada a natureza multifatorial e a heterogeneidade das doenças humanas, a validade dos modelos vem se tornando cada vez mais debatível (Gururajan et al., 2019; Seyhan, 2019; Ioannidis, 2016). Além disso, embora os animais tenham fornecido contribuições significativas para a compreensão e o avanço da ciência médica e biológica, há desafios referentes à translação dessas descobertas para humanos.

O termo translação tem sido aplicado para explicar a ponte que conecta a pesquisa básica à pesquisa clínica, visando a aplicação prática das descobertas da ciência básica para a prevenção e o tratamento de pacientes na clínica (Seyhan, 2019). Há cada vez mais debates de que vivemos uma crise de translacionalidade e de reprodutibilidade na pesquisa básica, referida na literatura como “*Valley of Death*”, onde os resultados falham ao serem aplicados em humanos por se tratar de dados não reprodutíveis e não confiáveis (Seyhan, 2019; Ioannidis, 2016; Chamuleau et al., 2018). A reprodutibilidade — ou capacidade de produzir resultados semelhantes por meio de estudos independentes — é essencial para a validade do método científico, mas estudos atuais indicam que ela é precária em quase todas as áreas de pesquisa básica (Voelkl et al., 2020), afetando a confiabilidade dos resultados. As falhas de translação e de reprodutibilidade podem ainda ser agravadas por erros metodológicos, fazendo-se necessário o aprimoramento do rigor científico através de medidas como cegamento, randomização, cálculo de tamanho amostral e relato adequado, inclusive os negativos ou sem relevância estatística (Chamuleau et al., 2018; Von Mücke-Heim et al., 2022), afinal, estima-se que apenas 11% de todos os estudos da pesquisa básica levam ao registro de intervenções terapêuticas (Chamuleau et al., 2018). Diante deste cenário, também há cada vez mais preocupações éticas sobre questões relativas ao bem-estar animal e a busca por alternativas ao uso de modelos animais está se tornando mais frequente na pesquisa científica.

O princípio dos 3Rs, expresso por W. M. S. Russell e R. L. Burch (1959), inspira ações para substituir, reduzir e refinar o uso de animais em pesquisas. De acordo com estes princípios, a redução se refere ao uso de uma menor quantidade de animais nos estudos sem que haja uma interferência significativa nas informações geradas; o refinamento se refere a esforços para reduzir a dor e o sofrimento dos animais durante os estudos; e a substituição se refere ao uso de métodos que não utilizem animais (Kaplan et al., 2024). Neste contexto, as revisões sistemáticas são abordagens importantes para o desenvolvimento de diretrizes em pesquisa, que podem contribuir com o rigor científico e podem ser usadas para identificar lacunas na literatura e orientar esforços em pesquisas, ou seja, podem sinalizar onde é necessário conduzir mais experimentos e onde é desnecessário gastar recursos, contribuindo para a redução do uso de animais em pesquisas (**Figura 1**) (Chamuleau et al., 2018; Ineichen et al., 2024; Ritskes-Hoitinga et al., 2019). Sendo assim, as revisões sistemáticas maximizam o valor dos estudos existentes e geram novos resultados sem experimentação animal adicional, alinhando-se ao princípio dos 3Rs, contudo, essa validade depende de serem revisões bem conduzidas e incluam estudos de qualidade.

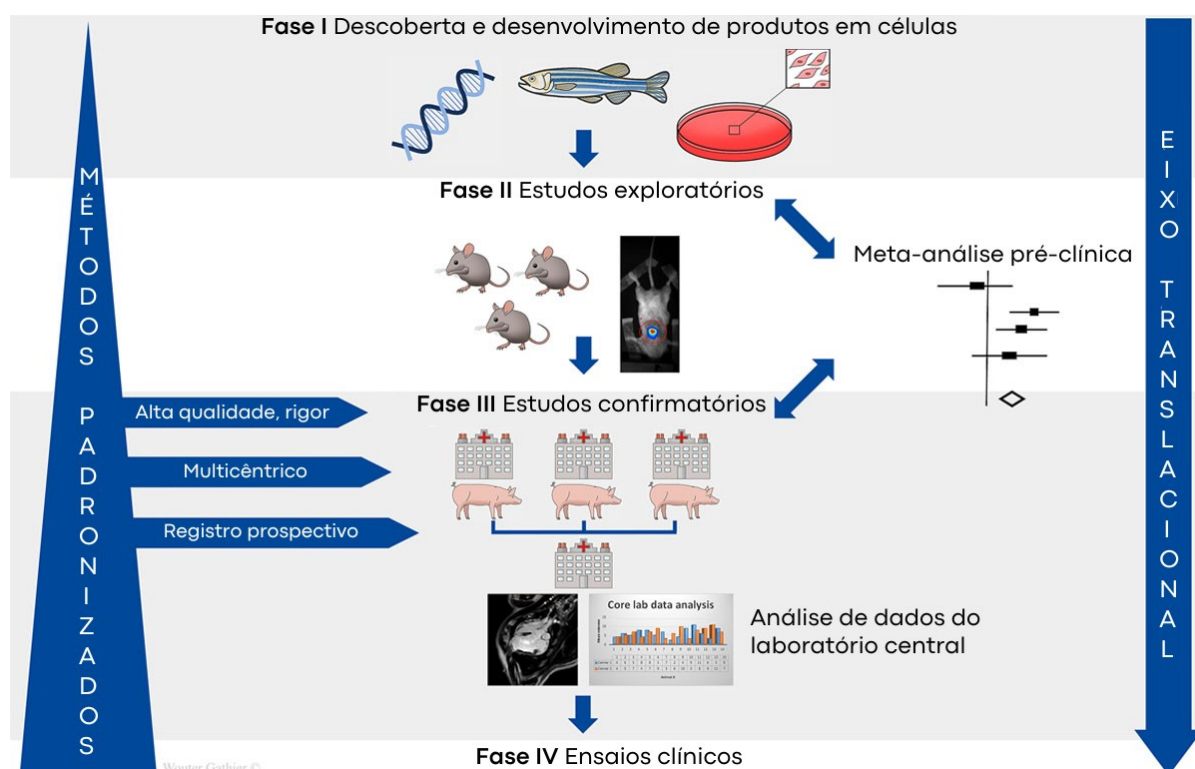


Figura 1. Estratégias de aprimoramento do eixo translacional em pesquisas pré-clínicas. Meta-análises pré-clínicas são realizadas com dados de estudos exploratórios e confirmatórios com o objetivo de fornecer uma visão geral dos dados disponíveis, apontar as lacunas e fornecer um direcionamento para futuras pesquisas pré-clínicas. A padronização dos métodos se intensifica conforme se avança ao longo do eixo translacional. *Reproduzido e traduzido de Chamuleau et al, 2018, com permissão de Wolters Kluwer Health, Inc.*

1.2 Revisões sistemáticas e meta-análises como alternativa à experimentação animal

As revisões sistemáticas e meta-análises (RSMA) são sínteses de evidências que utilizam métodos rigorosos e reproduzíveis para responder a uma pergunta de pesquisa específica. Em áreas como a medicina baseada em evidências, são consideradas as evidências de qualidade mais alta, atuando como uma lupa para as pesquisas primárias (**Figura 2**) (Ineichen et al., 2024; Murad et al. 2016). Diferenciam-se das revisões narrativas por terem uma metodologia explícita, por usarem ferramentas para avaliar e minimizar vieses e por serem conduzidas com ou sem uma meta-análise (Ineichen et al., 2024). Tais abordagens são úteis para fornecer conclusões mais confiáveis sobre, por exemplo, a relação causal entre uma intervenção e os efeitos observados, e são frequentemente empregadas para resolver controvérsias da literatura (Chamuleau et al., 2018; Deeks et al., 2024).

Na área clínica, as revisões sistemáticas focam na magnitude dos efeitos do tratamento, enquanto na área biomédica, incluindo a Farmacologia, buscam elucidar os mecanismos biológicos subjacentes, avaliando também dados de eficácia e segurança (Ineichen et al., 2024). O grupo Cochrane é responsável por diretrizes específicas voltadas à condução de revisões sistemáticas e meta-análises clínicas (Higgins et al., 2024), enquanto grupos como o SYRCLE e a CAMARADES atuam desenvolvendo, aplicando e disseminando a metodologia de revisões sistemáticas para estudos com animais (Ritskes-Hoitinga et al., 2014).

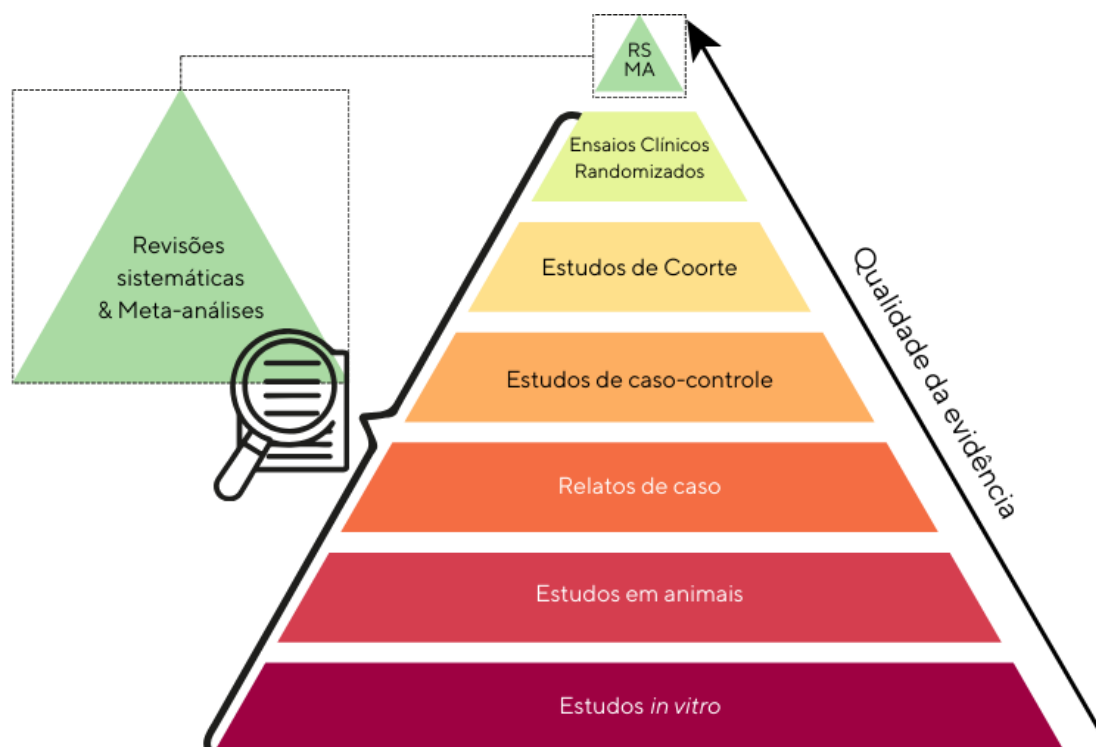


Figura 2. Pirâmide de evidência científica. Devido aos métodos padronizados e rigorosos, revisões sistemáticas

e meta-análises estão no topo da pirâmide de evidências, possuindo a mais alta qualidade e validade científica, quando bem conduzidas. São a síntese de todas as evidências científicas disponíveis, englobando toda a base da pirâmide, atuando como uma espécie de lupa para as pesquisas primárias. *Adaptado e traduzido de Murad et al. (2016) com permissão de BMJ Publishing Group Ltd*

A etapa inicial é o que define a revisão, onde se decide com precisão o que será investigado. Para isso, utiliza-se a estratégia PICO, que subdivide a pergunta em quatro componentes essenciais: a População (P) a ser estudada, a Intervenção (I) ou exposição analisada, o Controle (C) ou grupo de comparação e o Desfecho (O) (do inglês *outcome*), que representa os resultados a serem avaliados. Na fase de elaboração do protocolo, definem-se os métodos que evitarão vieses durante a execução da RSMA. As subetapas incluem a estratégia de busca (escolha de bases virtuais, palavras-chave e operadores booleanos), a estratégia de seleção (estabelecimento de critérios para triagem) e os métodos de extração de dados, que são divididos em qualitativos (focados na validade interna e externa) e quantitativos (métodos e síntese dos dados). Além disso, planeja-se a meta-análise (definindo modelo estatístico, análise de subgrupo, heterogeneidade e viés de publicação) e os meios de divulgação dos achados. A terceira etapa foca na transparência e no rigor científico, pois consiste no registro do protocolo em plataformas públicas, como o PROSPERO ou o *Open Science Framework* (OSF), com o objetivo de tornar o planejamento da revisão público antes do início da coleta de dados, o que previne a duplicidade de estudos e garante que os pesquisadores não alterem seus métodos conforme os resultados surjam. A etapa de implementação é a execução prática do protocolo estabelecido, envolvendo a realização das buscas nas bases virtuais, a seleção de estudos e a aplicação dos critérios de inclusão e exclusão. Em seguida, realiza-se a extração de dados (quantitativos e qualitativos) e a avaliação da validade interna (riscos de viés). O processo culmina na síntese qualitativa e na síntese quantitativa (meta-análise), onde são realizados os cálculos de tamanho do efeito (individual, global e de subgrupos) e as avaliações de heterogeneidade e viés de publicação. Após a análise dos dados, o foco se redireciona para a comunicação dos resultados. Esta etapa compreende o relato e divulgação dos resultados de forma clara e ética, seguindo diretrizes internacionais (como o PRISMA), culminando na publicação da revisão em periódicos da área de pesquisa. A etapa das revisões sistemáticas vivas representa o modelo mais atual de síntese de evidências, focado em continuidade. Envolve a atualização dos métodos com frequência e a definição da periodicidade para a repetição de todas as fases anteriores: novas buscas, seleção de novos estudos, extração de dados adicionais, reavaliação da qualidade e atualização das sínteses e meta-análises. Isso garante que a revisão reflita sempre a evidência mais atual disponível.



Figura 3. Etapas da condução de revisões sistemáticas e meta-análises. Em geral, as revisões sistemáticas e meta-análises possuem seis fases de desenvolvimento: elaboração da pergunta de pesquisa, elaboração do protocolo da revisão sistemática, registro do protocolo, implementação da pesquisa de revisão sistemática e meta-análise, publicação e atualização (revisões vivas). Cada etapa possui objetivos distintos que devem ser cumpridos para que a RSMA seja bem conduzida. *Adaptado de Bolzan & Oliveira (2022) sob licença CC BY-SA 4.0.*

Garantir o rigor metodológico na condução de uma RSMA tem sido um desafio crescente frente ao volume exponencial de dados na literatura científica (Marshall & Wallace, 2019). Estima-se, apenas para a área biomédica, que mais de um milhão de novos artigos sejam indexados anualmente na base do PubMed, criando um cenário onde o excesso de informação dificulta a triagem manual e a síntese de evidências (Landhuis, 2016; Bastian et al., 2010). Na literatura, alguns autores descrevem a produção científica atual como um “dilúvio de dados”, e

que isso exigiria uma transição da leitura manual tradicional para o que Ineichen, Rosso e Macleod (2023) denominam como “publômica” (*publomics*) — como em genômica e proteômica, onde “ômica” se refere à quantificação e análise de “um grande número de moléculas” (Perakakis et al., 2019). A discrepância entre a produção de evidências primárias e a síntese dessas evidências se acentuou significativamente nas últimas décadas. Enquanto os dados de Bastian et al. (2010) indicavam uma proporção de 75 ensaios clínicos para 11 revisões por dia, estimativas mais recentes mostram que o volume cresceu de forma acelerada. De acordo com bases de dados atualizadas em 2020 e 2021, os números indicam uma nova proporção de 142 ensaios clínicos para 80 revisões por dia (Marshall et al., 2020; Hoffmann et al., 2021). A consequência é um gargalo metodológico em que a maioria das revisões se encontra desatualizada e incapaz de incorporar a totalidade dos estudos relevantes (Marshall & Wallace, 2019). Calcula-se que, em média, 7% das RSMA já estão desatualizadas no ato de suas publicações (Shojania et al., 2007), com um tempo estimado de 11 meses para serem concluídas e 16 meses para serem publicadas (Bugajska et al., 2025).

Há uma sobrecarga de dados que os métodos manuais de revisão não conseguem mais acompanhar (Bastian et al., 2010), evidenciando a necessidade de se adotar métodos inovadores e mais eficientes para garantir a atualização e a aplicabilidade do conhecimento produzido. Em estudos pré-clínicos, embora a necessidade de evidências atualizadas não seja uma urgência como é na área clínica, a obsolescência das revisões é um problema. Como exemplo prático, uma revisão sistemática por Bolzan (2020) possuía uma biblioteca com 43 publicações até janeiro de 2020. Dois anos depois, uma atualização do banco de dados incluiu mais 23 publicações. A extração manual dos dados dessas referências adicionadas, necessária para manter as conclusões da revisão por Bolzan (2020) devidamente atualizadas, levou aproximadamente 12 semanas (Hamze, 2023). Isso demonstra que os problemas de sobrecarga de dados e obsolescência também ocorrem quando métodos manuais são aplicados em revisões na área da Farmacologia.

1.3 Automatização de revisões sistemáticas e meta-análises

O desenvolvimento tecnológico e computacional permitiu a criação de ferramentas que visam a otimização dos resultados das RSMAs e, conseqüentemente, a diminuição considerável do tempo gasto em cada etapa (Jonnalagadda, Goyal & Huffman, 2015). Contudo, historicamente houve um progresso desigual no desenvolvimento de ferramentas de automação para revisões sistemáticas. Enquanto as etapas de busca e de seleção de estudos atingiu um nível de maturidade considerável, a extração de dados permanece pouco desenvolvida (Marshall &

Wallace, 2019; Jonnalagadda et al., 2015). Ao contrário da etapa de triagem, em que as decisões são binárias (i.e., incluir/excluir), a extração de dados envolve a identificação de entidades e relações complexas entre palavras em textos não estruturados ou em tabelas, tornando o desenvolvimento de ferramentas universais muito mais desafiador (Marshall & Wallace, 2019).

A disponibilidade de *softwares* para gerenciamento de referências como o *Covidence* ou *Rayyan* ou SyRF auxiliam na seleção ou triagem dos estudos, acelerando esses processos e ajudando a mitigar discordâncias entre os revisores independentes (Veritas Health Innovation, 2014; Ouzzani et al., 2016; Babor et al., 2021). Além disso, quando se trata de triagem de títulos e resumos, classificadores de texto conseguem reduzir a carga de leitura em até 50% com alta sensibilidade (O'Mara-Eves et al., 2015). Bases bibliográficas como o *Medline* e o *Embase*, por exemplo, podem ser acessadas através de ferramentas de busca tais como o *Pubmed* e o *Ovid*, que auxiliam com estratégias de buscas avançadas de referências (NCBI, 1998; Wolters Kluwer, 1992).

Atualmente, ferramentas semiautomatizadas utilizando inteligência artificial já auxiliam também na extração de elementos PICO (População, Intervenção, Controle e Desfecho) e o aprendizado de máquina (*machine learning*, ML) têm sido cada vez mais considerada como uma solução prática tanto para a tarefa de extração quanto para triagem (Marshall et al., 2020). Classificadores de ML podem ser treinados para priorizar estudos relevantes durante a triagem de títulos e de resumos no processo de seleção de uma revisão sistemática, reduzindo a carga de trabalho manual em projetos de grande escala (Marshall & Wallace, 2019). O aprendizado de máquina também permite que sistemas aprendam padrões e façam previsões autonomamente a partir de dados, sem a necessidade de serem explicitamente programados para cada regra (Brown, 2021).

Marshall e Wallace (2019) argumentam que a automação não é mais uma promessa futurista, mas uma necessidade prática para lidar com o volume de evidências (Marshall & Wallace, 2019). Um estudo por Bannach-Brown et al. (2019), por exemplo, demonstrou que ao aplicar algoritmos de aprendizado supervisionado em uma revisão sistemática de estudos com animais foi observada uma redução de carga de trabalho de triagem superior a 90% em algumas configurações, mantendo uma sensibilidade de 95% a 99%. Os autores ressaltam que, além de poupar tempo, a automação também seria responsável por mitigar o erro humano inerente a tarefas repetitivas e exaustivas, como a triagem de milhares de referências, sugerindo que a máquina pode ser, em certos aspectos, mais confiável que o revisor humano (Bannach-Brown

et al., 2019). Adicionalmente, um estudo clínico de caso realizado por Clark et al. (2020) demonstrou a viabilidade de realizar uma revisão sistemática completa em apenas duas semanas. Ao utilizarem um conjunto de ferramentas de automação para triagem, obtenção de textos completos e avaliação de risco de viés, os pesquisadores conseguiram reduzir processos que manualmente levariam meses para serem concluídos, alegadamente sem comprometer a precisão metodológica (Clark et al, 2020).

Porém, apesar da existência de algumas ferramentas de *software* para RSMAs na área médica, a disponibilidade de ferramentas automatizadas para apoiar a condução de revisões sistemáticas na área biomédica são escassas, ao mesmo tempo que o número de artigos publicados tende a ser ainda maior no futuro (Bornmann & Mutz, 2015). O grupo SYRCLE admite que as RSMAs pré-clínicas precisam emprestar esses *softwares* por haver uma carência de ferramentas específicas que consigam lidar com a complexidade dos dados biológicos da experimentação animal (PROSPERO, 2026; Menon et al., 2021). Pesquisas pré-clínicas geram dados muito mais heterogêneos do que pesquisas clínicas, portanto, as ferramentas treinadas com dados clínicos frequentemente falham ao tentar extrair dados de estudos animais; além disso, essa falha é intensificada devido à falta de padronização nos relatos da pesquisa biomédica (Macleod et al., 2015; Kilkenny et al., 2010). Por exemplo, ferramentas de extração de dados como o *RobotReviewer*, baseada em ML, tiveram sucesso na extração automática de dados para análises de risco de viés em ensaios clínicos randomizados (Marshall et al., 2016), mas estudos clínicos apresentam perguntas de pesquisa mais específicas, desenhos experimentais mais estruturados e formas de publicação mais homogêneas que os ensaios pré-clínicos ou estudos em animais (Macleod et al, 2015; Kilkenny et al., 2010). Portanto, o *RobotReviewer* e demais ferramentas carecem da capacidade de extrair informações de textos que apresentem perguntas de pesquisa mais amplas ou nuances experimentais características dos estudos farmacológicos (Marshall et al., 2016; Soboczenski et al., 2019).

Um outro exemplo da aplicação de ferramentas baseadas em ML foi durante a pandemia do SARS-CoV-2, onde revisões sistemáticas de estudos clínicos foram aceleradas para identificar casos de pacientes infectados com diferentes cepas genéticas do vírus e para sugerir medidas para minimizar a transmissão a longo prazo (Garritty et al., 2021; Wang et al., 2021). A ferramenta COVID-NMA possibilitou a automatização de dados dos ensaios clínicos sobre a COVID-19, fornecendo um mapeamento vivo dos estudos (Boutron et al., 2020). Esse mapeamento forneceu uma síntese de evidências completa, atualizada e de alta qualidade sobre intervenções para a prevenção e o tratamento da COVID-19 (Davidson et al., 2023). Todavia,

sua aplicabilidade é restrita a ensaios clínicos e, para além disso, a ferramenta se encontra sem atualização desde 2023, quando a iniciativa foi concluída (The COVID-NMA Initiative, 2023). Nesse contexto, o desenvolvimento de outras ferramentas de extração de dados baseadas em tecnologias de inteligência artificial (IA) poderia transformar a síntese de evidências experimentais na biomedicina (Ineichen, Rosso & Macleod, 2023).

1.4 Inteligência artificial como ferramenta auxiliar para extração de dados

A IA foi consolidada por Russel e Norvig (2020) como um agente racional que é capaz de perceber seu ambiente e de agir sobre esse ambiente, agindo de modo a maximizar a probabilidade de alcançar resultados, independentemente de seu processo interno mimetizar o raciocínio humano ou não (Russell & Norvig, 2020). As abordagens iniciais no desenvolvimento da IA focavam na mimetização da cognição humana, como a do matemático britânico Alan Turing (1912-1954) que propôs que a atribuição de inteligência a uma máquina dependeria de seu sucesso em simular o comportamento humano a ponto de se tornar indistinguível dele, mas a perspectiva moderna se distanciou da necessidade de replicar processos biológicos (Russell & Norvig, 2020). Essa mudança se deu, em grande parte, devido à influência de contextos filosóficos como os apresentados por John Searle (1980), que argumentou que simular não é o mesmo que ser, redirecionando o foco da IA para a eficiência de resultados e não para a consciência da máquina (Searle, 1980). Sob essa ótica, a validade da IA na ciência reside na sua capacidade de processamento, e não na emulação da consciência.

Ao assumir a função de um agente racional focado em maximizar objetivos, como a extração de dados, a IA torna-se uma solução para enfrentar a demanda da “publômica”, pois se a máquina é capaz de escalar a análise de literatura a níveis cognitivamente inviáveis para pesquisadores humanos, é justamente essa capacidade de processamento que viabilizaria sua aplicação em tarefas repetitivas e intensivas (Ineichen, Rosso & Macleod, 2023; Marshall & Wallace, 2019). A capacidade da IA de resolver problemas complexos na ciência foi globalmente reconhecida quando Demis Hassabis e John Jumper, laureados com o Nobel de Química em 2024 juntamente com David Baker (Nobel Press, 2025), demonstraram a aplicabilidade da IA ao desenvolverem o AlphaFold2. Esse modelo abordou um problema de longa data da biologia estrutural: prever as estruturas tridimensionais complexas de proteínas apenas a partir de suas sequências lineares (Nature Portfolio, 2024), um feito que reflete o mesmo potencial de processamento de dados complexos que se busca aplicar na síntese de evidências científicas. Portanto, é válido ponderar o advento da inteligência artificial como uma tecnologia transformadora para implementar as etapas das RSMA's em áreas como a

Farmacologia (Elbadawi et al., 2024).

A etapa de extração de dados constitui uma das etapas mais enfadonhas, demoradas e propensas a erros de uma revisão sistemática. Estudos indicam que essa etapa consome uma parcela substancial dos recursos relacionados ao tempo despendido em um projeto, muitas vezes competindo apenas com a triagem inicial em termos de carga horária, mas a superando em complexidade cognitiva (Jonnalagadda et al., 2015). Além dessa sobrecarga de recursos, a ambiguidade na definição dos parâmetros de dados impacta severamente a concordância entre os avaliadores, ressaltando a importância de utilizar metadados claros e protocolos pré-definidos (Hoffmann et al., 2021). Outro aspecto importante é a inconsistência na apresentação dos resultados primários, especialmente de dados numéricos. Isso porque as variações de formato que os estudos relatam seus resultados (e.g., alternar entre médias e medianas ou desvios padrão e erros padrão da média) criam barreiras que dificultam a síntese quantitativa e qualitativa dos dados, exigindo intervenções manuais exaustivas para a conversão e a padronização desses valores (Hoffmann et al., 2021).

Neste contexto, o desenvolvimento tecnológico dos agentes de IA tem potencial de preencher a lacuna da extração de dados na síntese da literatura científica (Elbadawi et al., 2024). Embora inicialmente o uso ferramentas baseadas em inteligência artificial exigisse conhecimento em linguagem de programação (Ineichen et al., 2023), as interfaces atuais baseadas em Processamento de Linguagem Natural (*Natural Language Processing*, NLP) estão democratizando seu acesso, permitindo a replicação da fluidez da comunicação humana e interações sem a necessidade de comandos de programação complexos, que outrora representavam uma barreira para pesquisadores da saúde (Elbadawi et al., 2024).

Recentemente, a introdução dos Grandes Modelos de Linguagem (*Large Language Model*, LLM), como o GPT, revolucionou as possibilidades para aplicação em pesquisa científica (**Figura 4**). O uso desses modelos funciona como um “cérebro artificial”, capaz de interpretar instruções em uma linguagem natural para humanos a fim de realizar tarefas específicas (Qureshi et al., 2023). Os LLM são definidos como avançados sistemas de aprendizado profundo (*deep learning*, DL) que revolucionaram o NLP ao superar as limitações de redes neurais anteriores (Naveed et al., 2023). São baseados na arquitetura Transformer, apresentada no artigo “*Attention Is All You Need*” por Vaswani et al. (2017), considerado um divisor de águas no desenvolvimento da inteligência artificial, tornando possível a existência dos modelos atuais e servindo como a tecnologia base para os *chatbots* modernos (Diogenes,

2023; Zhao et al., 2023). De modo geral, o funcionamento dos LLM reside em mecanismos que permitem ao algoritmo ponderar a importância de diferentes palavras em uma frase, calculando o contexto e as relações entre os elementos sintáticos e semânticos (Naveed et al., 2023). Essa função é adquirida através do pré-treinamento em conjuntos massivos de dados textuais, onde o modelo aprende a modelar a distribuição estatística da linguagem para prever, por probabilidade estatística, o próximo *token* (unidade de texto) em uma sequência.

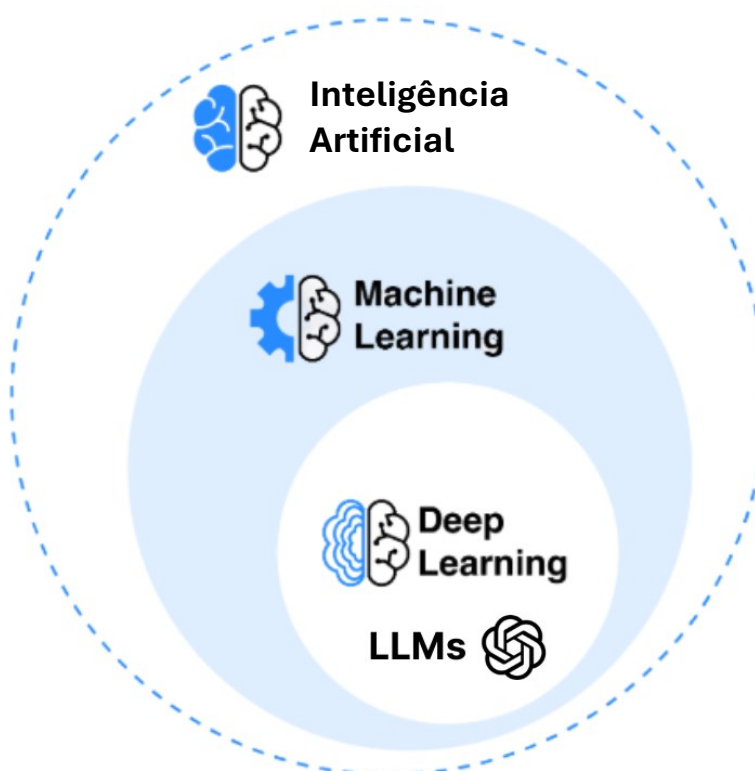


Figura 4. Hierarquia dos domínios da inteligência artificial. O diagrama apresenta o aprendizado de máquina (*machine learning*) como um subconjunto da inteligência artificial (IA), que por sua vez engloba o aprendizado profundo (*deep learning*). Os Grandes Modelos de Linguagem (LLMs), representados pelo ícone do ChatGPT, constituem a camada mais específica desta especialização tecnológica. *Adaptado e reproduzido de NASA (2024).*

Tecnicamente, os *chatbots* de IA representam a interface de usuário para sistemas complexos que simulam conversas humanas (**Figura 5**). Esses sistemas analisam a entrada (*input*), como o texto completo de um artigo científico, e o interpretam para gerar uma resposta apropriada (*output*), como uma tabela de dados estruturada. Isso ocorre pela combinação das tecnologias dos LLM, i.e., Processamento de Linguagem Natural e Aprendizado de Máquina, onde o NLP permite que a máquina interprete a sintaxe e a semântica da linguagem biomédica, enquanto os algoritmos de ML permitem o refinamento do desempenho (Zhao et al., 2023). A vantagem dos LLM é a capacidade de aprendizado onde o modelo pode realizar uma tarefa de extração de dados complexa sem ter sido especificamente treinado para aquele tópico, apenas

seguindo as instruções do *prompt* fornecido pelo usuário (Wang et al., 2024). Isso viabiliza sua adaptação para tarefas específicas, como a extração de dados na literatura científica.

Um estudo de caso utilizando LLM para extração de dados em síntese de evidências descobriu que tais modelos podem extrair características de estudos e resultados com uma precisão comparável à de revisores humanos sem experiência, embora exijam verificação para dados altamente complexos (Gartlehner et al., 2024). Outro estudo apresentou resultados semelhantes ao indicar que a automação de revisões sistemáticas com LLM atingiu um patamar de alta fidelidade com a triagem e a extração de dados, sugerindo que estamos nos aproximando de um momento em que a IA poderá assumir a maior parte da carga de trabalho observada na condução de uma RSMA (Cao et al., 2025).

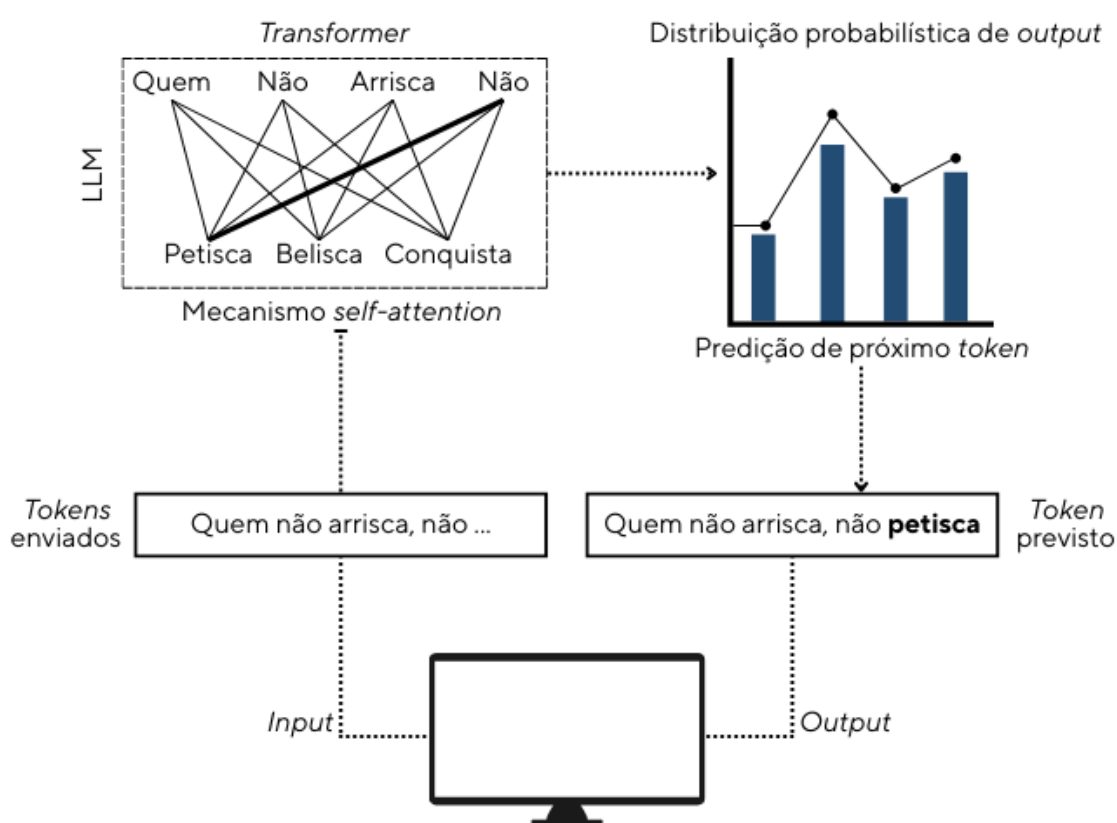


Figura 5. Representação esquemática do processo de geração de texto em um LLM. O diagrama ilustra como a arquitetura *Transformer* utiliza o mecanismo de *self-attention* para analisar o contexto do *input*. O modelo calcula uma distribuição probabilística entre possíveis candidatos e seleciona o *token* com maior probabilidade estatística para compor o *output*. Adaptado de Naveed et al. (2023) sob licença CC BY 4.0.

Entretanto, a aplicação direta dos LLM na extração de dados biomédicos exige cautela metodológica, contraponto o entusiasmo inicial à luz de evidências sobre as limitações de desempenho em tarefas de alta especificidade (Alshami et al., 2023; Gutierrez et al., 2022; Messeri & Crockett, 2024). Gutierrez et al. (2022) demonstrou que modelos generalistas que

operam apenas sob a lógica de “aprendizado no contexto”, onde o modelo recebe apenas alguns exemplos no *prompt* sem ser re-treinado, podem apresentar desempenho inferior a modelos menores, porém especializados e submetidos a ajuste fino, como o PubMedBERT. Os autores observaram que o GPT-3 apresentou dificuldades em manter a precisão rigorosa exigida no domínio biomédico (Gutierrez et al., 2022). Complementarmente, Alshami et al. (2023) exploraram o uso do ChatGPT em todas as etapas da revisão sistemática, desde a formulação da estratégia de busca até a redação do manuscrito. Embora a ferramenta tenha acelerado significativamente o processo e auxiliado na geração de estruturas lógicas e códigos de busca, como observado nos estudos citados anteriormente, os autores alertam para limitações críticas, como a tendência à “alucinação”, isto é, geração de informações falsas ou referências inexistentes. De fato, devido à natureza puramente probabilística dos LLM, desprovida de compreensão de significado, linguistas como Bender et al. (2021) alertam para o uso de tais modelos, caracterizando-os como “papagaios estocásticos”.

Assim, de acordo com a literatura, o treinamento massivo do modelo (GPT) não garante, por si só, a acurácia na estruturação de dados complexos, indicando que a substituição de *pipelines* especializados por *chatbots* generalistas não é uma escolha trivial e pode introduzir erros se não houver estratégias de adaptação específicas para a área médica ou biomédica, além de uma validação rigorosa por avaliadores humanos (Gutierrez et al., 2022; Alshami et al., 2023; Messeri & Crockett, 2024). Em vista disso, a metodologia mais adequada envolveria uma hibridização, onde a IA atuaria como um auxiliar sob estrita supervisão humana.

Portanto, para averiguar se há meios de substituir parcialmente o avaliador humano, aplicamos modelos de IA como ferramenta auxiliar na etapa de extração de dados de uma revisão sistemática sobre a teoria neurogênica da depressão e investigamos a viabilidade desta substituição, isto é, se a inteligência artificial foi capaz de realizar a extração com a mesma acurácia que um avaliador humano, e se os dados extraídos são confiáveis o suficiente para a síntese. A partir disto, buscamos compreender se as ferramentas baseadas em IA podem, de fato, romper o gargalo histórico da extração de dados, otimizando o tempo na condução de revisões sistemáticas, contribuindo para a validação de novos métodos na síntese de evidências.

2 HIPÓTESE

A aplicação de um protocolo computacional baseado em Grandes Modelos de Linguagem (LLM) permite a extração de dados pré-clínicos com acurácia equivalente ao avaliador humano e com eficiência temporal significativamente superior.

3 OBJETIVO GERAL

Desenvolver e validar um protocolo computacional baseado em inteligência artificial para automação da etapa de extração de dados em revisões sistemáticas de estudos pré-clínicos, avaliando sua viabilidade técnica e acurácia por meio de um estudo piloto.

3.1 Objetivos específicos

a) Padronização:

- i. Desenvolver e estruturar um protocolo de extração automatizada utilizando Grandes Modelos de Linguagem (LLMs) para variáveis farmacológicas;
- ii. Demonstrar a aplicabilidade do protocolo de extração padronizado por meio de um fluxo de trabalho em uma amostra representativa de uma revisão sistemática real para extrair dados quantitativos e qualitativos em um LLM (ChatGPT);
- iii. Avaliar a transferibilidade aplicando o protocolo de extração padronizado a um LLM distinto (Gemini).

b) Validação:

- i. Mensurar o desempenho temporal da extração automatizada em comparação à extração manual;
- ii. Avaliar a acurácia da extração automatizada em relação ao padrão-ouro (extração manual) por meio de análises de concordância e métricas de desempenho;
- iii. Comparar o desempenho das extrações pelo ChatGPT e Gemini, verificando a precisão na identificação dos parâmetros experimentais e farmacológicos.

4 METODOLOGIA

Este estudo se baseou no desenvolvimento metodológico e validação conceitual de um protocolo computacional para automação da extração utilizando dados de uma revisão sistemática sobre neurogênese hipocampal induzida por antidepressivos como prova de conceito. A amostra de 26 artigos em formato PDF que foram utilizados para o desenvolvimento deste estudo pertence à atualização do banco de dados de uma revisão sistemática por Bolzan (2020), cujos dados foram extraídos manualmente por dois revisores independentes e conciliados por um terceiro revisor (Hamze, 2023; Bolzan, 2020). Esse banco de dados atualizado foi considerado o padrão-ouro ou a referência de acerto para a validação da acurácia da extração automatizada.

A metodologia foi estruturada de acordo com os objetivos específicos, conforme descrito a seguir: 1) padronização e desenvolvimento do protocolo de extração; 2) aplicabilidade do protocolo de extração em uma amostra representativa; 3) avaliação da transferibilidade; e 4) validação da acurácia e eficiência da extração automatizada.

4.1 Computador, *softwares* e pacotes utilizados

A acessibilidade e a reprodutibilidade do protocolo de extração foram demonstradas pela execução integral das etapas em ambiente local, utilizando um computador pessoal de especificações básicas (processador Intel Core i3 2.3 GHz, 4GB de RAM e SSD de 240GB), dispensando o uso de infraestrutura de nuvem ou máquinas virtuais e reforçando a viabilidade do fluxo de trabalho sem a necessidade de recursos computacionais de alto desempenho.

A padronização e validação dos dados foram realizadas no ambiente do RStudio (Posit Team, 2025), integrando as linguagens R (versão 4.4.3) (R Core Team, 2025) e Python (versão 3.13) (Python Software Foundation, 2025) por meio do pacote *reticulate*. Para o tratamento de dados, empregaram-se os pacotes *readxl*, *tidyr* e *dplyr* (R) e *pandas* (Python). O alinhamento dos dados e as análises estatísticas foram suportados pelas bibliotecas *scipy* e *scikit-learn* (Python). A visualização dos resultados foi executada com os pacotes *ggplot2*, *viridis*, *tibble*, *purrr* e *knitr* (R), complementados por *seaborn* e *matplotlib* (Python) para a geração das matrizes de confusão. A integridade das funções de tratamento foi validada por meio de testes com a biblioteca *unittest* (Python).

4.2 Padronização e desenvolvimento do protocolo de extração

A etapa de padronização foi delineada para estabelecer um protocolo metodológico capaz de converter a tarefa de extração de dados em um fluxo de trabalho semiautomatizado e reprodutível. Esta etapa se dedicou à estruturação dos alicerces operacionais do protocolo, abrangendo desde a escolha dos modelos de IA até a construção e refinamento do comando de instrução (*prompt*). Nesta etapa, também foram definidos parâmetros de saída de dados para minimizar a variabilidade estocástica inerente aos modelos de linguagem (Zhao et al., 2023), assegurando que a extração automatizada seguisse diretrizes equiparáveis às do protocolo seguido por avaliadores humanos antes de ser submetida à validação. Para selecionar o modelo mais adequado para a extração de dados, foi realizada uma avaliação qualitativa dentre os LLM disponíveis (item 4.2.1). Para a execução da primeira etapa do estudo piloto, também foi necessária a realização da engenharia de *prompt* (item 4.2.2).

4.2.1 Seleção do modelo LLM

O principal critério para a seleção do LLM foi a disponibilidade de interfaces de conversação (*chatbots*) acessíveis a pesquisadores sem experiência com linguagem de programação, dispensando a construção de *scripts* de requisição via API (*Application Programming Interface*, que permite “emprestar” a inteligência dos modelos para uso localmente). Além disso, foram considerados critérios como: a capacidade de processar o texto dos artigos em uma única interação (janela de contexto) e a disponibilidade de planos que viabilizassem a reprodução do protocolo por terceiros. Todas as informações relativas às características dos modelos foram coletadas nos *websites* próprios de cada desenvolvedor, i.e., [OpenAI](#) (GPT), [Meta](#) (Llama), [Google DeepMind](#) (Gemini) e [Anthropic](#) (Claude).

4.2.2 Engenharia de *prompt*

Uma vez que os modelos do estudo piloto foram selecionados, iniciou-se a etapa de engenharia de *prompt*, que pode ser definida como o desenvolvimento iterativo das instruções de entrada (*input*). O *prompt* foi elaborado seguindo guias técnicos disponibilizados pela Anthropic na [Claude Developer Platform](#) (Anthropic, 2025). A estrutura do *prompt* foi desenhada integrando a ferramenta mnemônica PICO (P = população ou problema; I = intervenção ou tratamento; C = controle ou comparador; O = *outcome* ou desfecho ou resultado) com diretrizes estritas para fornecer respostas concisas e baseadas exclusivamente nos artigos fornecidos (Higgins et al., 2024). A construção se baseou em etapas iterativas de refinamento repetitivas até que a ferramenta buscasse, extraísse e tabulasse os dados de interesse de uma amostra dos artigos científicos conhecidos e com dados previamente extraídos (n=26).

O processo de construção do *prompt* se encerrou quando o número de dados extraídos se estabilizou, i.e., a informação retornada não se tornava maior ou mais complexa a cada nova interação. A engenharia de *prompt* foi desenhada para incluir regras específicas para a definição de estudos individuais, ou seja, separar múltiplos estudos dentro de um mesmo artigo em linhas distintas, considerando variações no desenho experimental. Também foram integradas táticas de controle, tais como a atribuição de *persona*, contextualizando o modelo a operar sob o viés analítico de um revisor sistemático, seguindo as boas práticas descritas nos manuais da Anthropic. Uma vez consolidado no ChatGPT, o mesmo comando de instrução foi aplicado ao modelo Gemini 2.5 Pro. Ao contrário da operação iterativa adotada para o ChatGPT, a interação no Gemini utilizou as instruções já refinadas. Com isso, verificamos que o *prompt* pôde ser aplicado em outros modelos sem necessidade de uma nova engenharia de *prompt*, bem como verificamos o desempenho dos LLMs distintos.

4.3 Aplicabilidade do protocolo de extração em uma amostra representativa

Para averiguar a aplicabilidade e o desempenho do protocolo de extração automatizada em um cenário real de revisão sistemática, esta etapa visou transpor o *prompt* para a execução prática, instruindo o modelo de linguagem a processar as informações específicas da farmacologia. O foco desta fase foi garantir que a IA operasse dentro dos limites dos dados biológicos fornecidos, visando minimizar a ocorrência de “alucinações” (geração de dados inexistentes) e assegurar que cada variável extraída mantivesse a fidelidade aos achados experimentais originais. Para isso, a aplicabilidade foi testada em interface de *chat*, simulando as condições reais de um pesquisador acadêmico da farmacologia sem experiência em código. Detalhes de como a extração foi conduzida são encontrados nas seções 4.3.1 e 4.3.2.

4.3.1 Características da amostra

Visto que a amostra proveio de uma revisão sistemática concluída, o protocolo original serviu de referência para a extração assistida por inteligência artificial. Portanto, foram extraídas as seguintes variáveis textuais ou numéricas: autor, ano, estudo, marcador de neurogênese, espécie, sexo, linhagem, antidepressivo, dados do antidepressivo (dose, regime de injeções, duração do tratamento), dados do marcador BrdU (dose, número de injeções e tempo de sobrevivência), protocolo de estresse e teste comportamental.

4.3.2 Extração automatizada e protocolo de execução

O plano gratuito do [ChatGPT](#) na versão utilizada continha uma restrição de *upload* e de leitura de arquivos, com limite de 10 arquivos a cada 24 horas. Por conta disso, o processamento do número total de artigos (n=26) foi dividido em três etapas ou ciclos de extração (n=10, n=10 e n=6). Já para o [Gemini](#), o acesso à versão 2.5 *Pro* foi possível graças a uma licença educacional disponibilizada gratuitamente aos estudantes de universidades ao redor do mundo, oferecendo assim uma maior capacidade de processamento e sem as restrições de *upload* e leitura. Dessa forma, a extração de dados foi realizada em um único ciclo. Para as planilhas geradas pelos modelos, foi necessário um pré-processamento dos dados antes da importação para o ambiente do RStudio, onde foram executadas ações para agregar as planilhas em um único arquivo, para remover formatações (i.e., cores, estilos de fontes, mesclagem de células e filtros) e para padronizar o cabeçalho das colunas.

4.4 Validação da acurácia e eficiência da extração automatizada

O protocolo de validação foi delineado para mensurar a validade técnica da ferramenta de IA em relação à extração manual, estabelecida como padrão-ouro. Inicialmente, a aplicação

direta do fluxo de trabalho, isto é, da extração à análise estatística, mostrou-se inviável devido à heterogeneidade terminológica e aos desafios na agregação dos dados brutos. Para mitigar essas inconsistências e assegurar a integridade da validação, procedeu-se ao tratamento e alinhamento dos dados via *scripts* em R e Python. Este processo compreendeu o tratamento dos dados (item 4.4.1), seguido pelo pareamento entre as extrações manual e automatizada (item 4.4.2). Subsequentemente, realizaram-se as análises estatísticas e a avaliação das métricas de desempenho para o modelo ChatGPT (item 4.4.3), metodologia esta replicada para o Gemini, permitindo uma análise comparativa rigorosa entre ambas as ferramentas (item 4.4.4).

4.4.1 Tratamento e padronização de dados

Para o tratamento dos dados, desenvolveu-se um fluxo de trabalho (*pipeline*) híbrido integrando as linguagens R e Python no ambiente RStudio (**Quadro 1**), por meio do pacote *reticulate* (Ushey et al., 2025). Essa abordagem permitiu combinar a robustez estatística da coleção *tidyverse* em R (Wickham et al., 2019) com a versatilidade das bibliotecas *re* e *pandas* (The Pandas Development Team, 2025) em Python, fundamentais para a manipulação de grandes volumes de dados e processamento de expressões regulares (*regex*). As funções de tratamento foram aplicadas de forma padronizada tanto ao conjunto de dados extraídos manualmente quanto aos conjuntos gerados pelos LLMs (ChatGPT e Gemini) e são descritas abaixo.

Para células contendo múltiplos dados (e.g., três doses diferentes em uma mesma célula), implementou-se uma correção de desagregação utilizando a função *explode* da biblioteca *pandas*. O processo se baseou em *regex* para detectar separadores textuais (como vírgula e ponto-e-vírgula) e desmembrar os agrupamentos. Esse procedimento garantiu que a unidade amostral fosse padronizada no formato de um experimento por linha, seguindo os princípios de organização de dados científicos em uma RSMA, permitindo melhor pareamento entre os métodos manual e automatizado. Sequencialmente ao procedimento de desagregação, o *script* de limpeza executou a normalização de formatação. Foi feita a conversão de caracteres para caixa baixa e foram aplicadas funções globais de limpeza para remover elementos textuais que não constituem dados (i.e., espaços e caracteres textuais) e que impedem a correspondência exata de *strings* (sequência de caracteres que forma um texto).

Essa etapa também empregou dicionários de mapeamento e lógica condicional para a normalização de variações terminológicas, transformando-as em siglas (e.g., “fluoxetine” para “flu”) (**Figura 4**). No caso das variáveis temporais, implementou-se uma função de conversão

capaz de identificar distintas unidades de tempo (e.g., horas, dias, semanas e meses) e convertê-las para um valor numérico padronizado em dias. O algoritmo de normalização de unidades incluiu regras de exceção para preservar a precisão em frações de horas, garantindo a integridade dos dados de curta duração quando pertinente ao protocolo experimental.

A padronização de espécies e linhagens de ratos e camundongos consistiu no mapeamento de variações linguísticas, seguido da unificação de sinônimos em termos padronizados (e.g., “mouse” para “mice”, “c57bl/6n” para “c57bl/6”, “rats” para “rat”). Esse procedimento foi igualmente aplicado aos protocolos de estresse e testes comportamentais, empregando-se um dicionário de redução para siglas universais (e.g., FST, TST, CUMS), permitindo que denominações distintas para um mesmo parâmetro fossem computadas de forma equivalente. Adicionalmente, para assegurar que o processamento não introduzisse distorções aos dados brutos, implementou-se um módulo de validação por meio de testes unitários com a biblioteca nativa *unittest* da linguagem Python.

```
py_run_string(r"(
import pandas as pd
import re

### Funções de Padronização de Texto ###
def remover_parentheses(df):
    '''Itera sobre todas as colunas de texto e remove o conteúdo entre parênteses.'''
    for col in df.select_dtypes(include=['object']).columns:
        df[col] = df[col].str.replace(r'\s*\([^)]*\)\s*', '', regex=True).str.strip()
    return df

def padronizar_minusculas(df):
    '''Converte qualquer texto minúsculas.'''
    df.columns = df.columns.map(str).str.lower().str.replace(' ', '_').str.replace('_(atd)', '_atd')
    for col in df.select_dtypes(include=['object']).columns:
        df[col] = df[col].str.lower().str.strip()
    return df

[...]
```

```
### Funções de Padronizações Específicas ###
def padronizar_especies(valor):
    '''Padroniza os nomes das espécies (ex: rats = rat, mouse = mice).'''
    if pd.isna(valor):
        return valor
    val_limpo = str(valor).lower().strip()
    if val_limpo in ['nan', 'na', 'n/a', 'nr']:
        return valor
    species_map = {
        'rats': 'rat',
        'mouse': 'mice'
    }
    return species_map.get(val_limpo, val_limpo)

def padronizar_strain(valor):
    '''Padroniza os nomes das linhagens (ex: sprague dawley = sprague-dawley).'''
    if pd.isna(valor):
        return valor
    val_limpo = str(valor).strip().lower()
    if val_limpo in ['nan', 'na', 'n/a', 'nr']:
        return valor
    strain_map = {
```

```

        'wistar kyoto': 'wistar-kyoto',
        'sprague dawley': 'sprague-dawley',
        'sprague-dawley': 'sprague-dawley',
        'camkiv null mice': 'camkiv',
        'camkiv null': 'camkiv',
        'balb/cj': 'balb/c',
        'c57bl/6n': 'c57bl/6',
        'swiss albino': 'swiss',
        '129/svj': '129s1',
        '129s1/svimj': '129s1',
        'lr line': 'ir line',
        'c57/bl6': 'c57bl/6'
    }
    return strain_map.get(val_limpo, val_limpo)

def padronizar_dose(valor):
    '''Padroniza a unidade de dose (ex: 10 = 10 mg/kg).'''
    if pd.isna(valor):
        return valor
    s_valor = str(valor).strip().lower().replace(' ', '')
    if s_valor in ['nan', 'na', 'n/a', 'nr']:
        return valor
    match = re.match(r'^(\d+(?:\.\d+)?)([a-zµ]*(?:/[a-zµ]+)?)', s_valor)
    if not match: return valor
    numero, unidade = match.groups()
    if unidade in ['', 'mg/kg', 'mgkg']: return f'{numero} mg/kg'
    else: return f'{numero} {unidade}'

[...]
```

```

antidepressants_map = {'fluoxetine': 'flx', 'imipramine': 'imi', 'amitriptyline': 'ami', 'citalopram':
'cit', 'desipramine': 'des', 'escitalopram': 'esc', 'paroxetine': 'prx', 'fluvoxamine': 'flv'}
antidepressants_map_otimizado = criar_mapa_otimizado(antidepressants_map)

testes_map = {'forced swim test': 'fst', 'forced swim': 'fst', 'porsolt test': 'fst', 'forced
swimming': 'fst', 'forced swimming test': 'fst', 'tail suspension test': 'tst', 'tail suspension':
'tst', 'sucrose preference test': 'spt', 'sucrose preference': 'spt', 'open field test': 'oft', 'open
field': 'oft', 'locomotor activity': 'oft', 'novelty-suppressed feeding test': 'nsf', 'novelty-
suppressed feeding': 'nsf', 'novelty-suppressed feeding (nsf)': 'nsf', 'contextual fear conditioning':
'fcp', 'fear conditioning paradigm': 'fcp', 'social interaction test': 'sit', 'social interaction':
'sit', 'splash test': 'spl', 'splash': 'spl', 'elevated plus maze': 'epm', 'nest test': 'nest',
'shuttle-box test': 'sbx', 'shuttle-box': 'sbx', 'morris water maze': 'mwm', 'water maze': 'mwm',
'unconditioning test with tone': 'utt', 'grip strength': 'gst', 'beam-walking': 'bwt', 'other
behavioral tests': 'other', 'fear conditioning': 'fcp', 'shuttle-box escape test': 'sbx', 'nest
building test': 'nest', 'rotarod': 'rot', 'actimeter': 'act', 'cylinder': 'cyl', 'morris water-maze
test': 'mwm', 'elevated plus maze test': 'epm', 'forced swim task': 'fst', 'tail suspension task':
'tst', 'beam-walking test': 'bwt', 'cued fear-conditioning test': 'fcp', 'cylinder test': 'cyl', 'grip
strength test': 'gst', 'open-field behavior': 'oft', 'rotarod test': 'rot', 'spontaneous locomotor
activity': 'oft', 'water maze test': 'mwm', 'novel suppressed feeding test': 'nsf', 'sucrose intake':
'spt'}
testes_map_otimizado = criar_mapa_otimizado(testes_map)

[...]
```

```

### Função para Aplicar Tratamento ###
def tratar_dataframe(df):
    df = df.copy()
    df = remover_parentheses(df)
    df = padronizar_minusculas(df)

    cols_to_exclude = ['id', 'author', 'year', 'study', 'stress_protocol', 'behavioral_test']
    cols_to_explode = [col for col in df.columns if col not in cols_to_exclude]

[...]
```

Quadro 1. Código desenvolvido para o tratamento de dados. Partes do código de tratamento demonstrando algumas das padronizações e regras aplicadas aos *dataframes* para garantir a correspondência exata de termos extraídos manualmente e de modo automatizado pelos modelos GPT e Gemini. *Elaborado pela autora (2025).*

4.4.2 Alinhamento e pareamento de estudos

O alinhamento e pareamento dos dados foi realizado através da implementação de um

algoritmo de melhor correspondência (*best-fit*), modelado matematicamente como um problema de atribuição linear (Christen, 2012). Implementou-se um algoritmo em Python utilizando a biblioteca *scipy* (Virtanen et al., 2020), fundamentado na lógica de maximização da similaridade global por artigo (ID). Para cada ID, estruturou-se uma matriz de custo comparando as extrações manual (FCH) e automatizada (ChatGPT ou Gemini) (**Figura 6**). Para cada célula da matriz foi atribuído um *score* de similaridade, calculado pela soma de igualdades exatas entre as variáveis de resultado. A resolução da matriz para identificar os pares com maior correspondência foi realizada por meio da função *linear_sum_assignment* (Virtanen et al., 2020), técnica consolidada para a resolução de problemas de atribuição linear em processos de integração de dados (Christen, 2012). O algoritmo de alinhamento também foi empregado para identificar e segregar registros sem correspondência direta entre os métodos. Informações presentes na extração manual (padrão-ouro), mas ausentes no conjunto gerado pela IA, foram classificadas como falsos negativos, indicando omissão de dados pelo modelo. Inversamente, dados extraídos pela IA que não possuíam correspondência na extração manual foram classificados como falsos positivos, indicando alucinações do modelo. Essa categorização foi essencial para o cálculo posterior das análises de concordância e das métricas de desempenho.

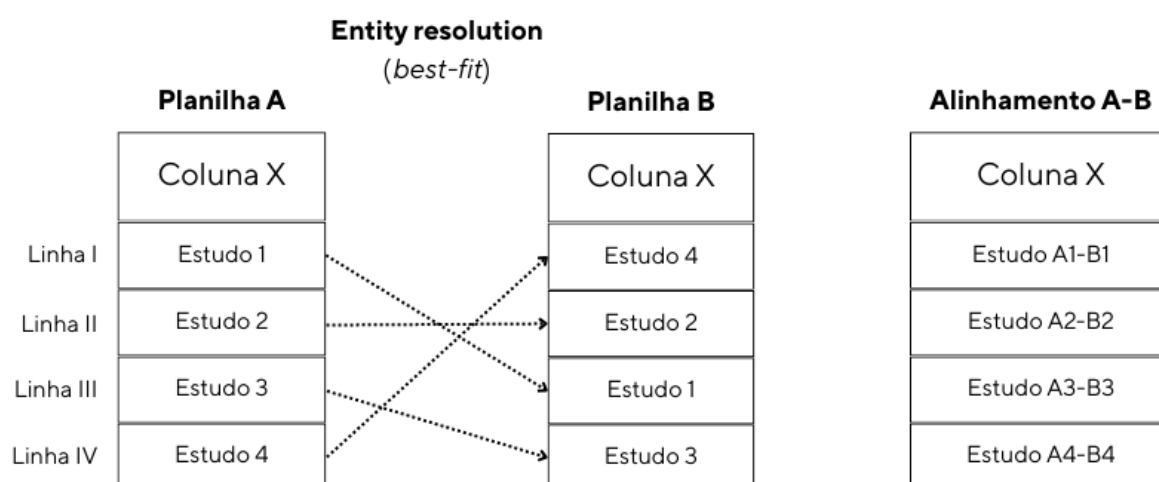


Figura 6. Mecanismo de resolução de entidades (*entity resolution*). Ilustração do método *best-fit* para o pareamento de estudos entre planilhas de extração e a geração de um alinhamento final de dados. O diagrama ilustra o pareamento de dados provenientes de diferentes fontes (Planilhas A e B), onde as setas indicam o reconhecimento de estudos correspondentes em diferentes ordens de catalogação (linhas), resultando em um alinhamento final (A-B) consolidado que permite a análise comparativa entre os estudos incluídos. *Elaborado pela autora (2025).*

4.4.3 Análises de concordância e métricas de desempenho

Todos os cálculos estatísticos foram processados por meio da biblioteca *scikit-learn* do Python (Pedregosa et al., 2011), enquanto as representações gráficas foram desenvolvidas com

os pacotes *ggplot2* do R (Wickham, 2016) e *seaborn* do Python (Waskom, 2021). As análises foram estruturadas em dois níveis complementares para garantir o rigor da validação. A nível de análise do conteúdo extraído por artigo (“nível de artigo”), utilizou-se o conjunto de dados **tratados** para mensurar a capacidade de recuperação de termos (*tokens*) e avaliar se a IA foi capaz de identificar as informações corretas dentro de cada artigo, independentemente da organização das linhas. Já a nível de análise de concordância por estudo (“nível de estudo”), as métricas foram aplicadas sobre o conjunto de dados **alinhados** após o pareamento pelo algoritmo *best-fit*. Este procedimento garantiu que a comparação ocorresse entre unidades experimentais equivalentes. Neste nível, empregaram-se métricas de desempenho utilizadas para avaliar a precisão da classificação e a concordância entre os métodos manual e automatizado para cada variável específica. Resumidamente, para a avaliação do protocolo, o processamento estatístico foi organizado em quatro etapas analíticas: 1) métricas de recuperação de informação a nível de artigo; 2) métricas de concordância de classificação a nível de estudo; e 3) integridade estrutural a níveis de artigo e de estudo.

As métricas de recuperação de informação (**Figura 6**) foram aplicadas a todas as colunas, tratando cada célula como um conjunto de *tokens* processados sob a abordagem *bag-of-words*, fundamentada na decomposição textual e identificação de termos individuais (Scikit-Learn Developers, 2025). Enquanto o protocolo de tratamento garantiu a normalização de termos farmacológicos e metodológicos em siglas uniformes (e.g., “Forced Swim Test” e “Porsolt’s Test” convertidos para “fst”), a lógica de conjuntos de *tokens* na fase de validação assegurou a invariância em relação à ordem estrutural. Dessa forma, a validação de acurácia não foi impactada por variações na sequência dos termos entre as planilhas manual e automatizada, priorizando a presença e a exatidão do conteúdo extraído.

O desempenho da extração foi quantificado pelas métricas de precisão (*precision*), sensibilidade ou revocação (*recall*) e F_1 -score (Powers, 2011). A precisão atesta a confiabilidade da ferramenta, indicando a proporção de dados extraídos que são fidedignos ao padrão-ouro, mitigando a ocorrência de inserções indevidas ou alucinações (falsos positivos). Paralelamente, a revocação avalia a sensibilidade do modelo em recuperar a totalidade das informações pertinentes, identificando eventuais omissões de dados (falsos negativos). Como indicador síntese de eficácia, utilizou-se o F_1 -score, definido pela média harmônica entre as métricas anteriores. Esta escolha justifica-se pela natureza desbalanceada dos dados textuais em artigos científicos e pela necessidade de um equilíbrio entre fidedignidade e abrangência, assegurando que o protocolo de validação penalize tanto a geração de informações inexistentes

quanto a perda de dados relevantes (Powers, 2011), refletindo com maior fidedignidade a utilidade prática da ferramenta para a extração de dados na farmacologia.

Precisão (P):

$$P = \frac{VP}{(VP + FP)}$$

Sensibilidade (R):

$$R = \frac{VP}{(VP + FN)}$$

F_1 -score (F_1):

$$F_1 = 2 \times \frac{(P \times R)}{(P + R)}$$

Em que:

- VP : verdadeiro positivo;
- FP : falso positivo (alucinação);
- FN : falso negativo (omissão).

Para as variáveis extraídas, foram calculadas métricas de desempenho baseadas na teoria de conjuntos: para cada par de artigo (Manual vs. ChatGPT ou Gemini), definiu-se que os verdadeiros positivos são *tokens* presentes na interseção entre o conjunto manual e o conjunto predito pelo ChatGPT ou Gemini; os falsos positivos são *tokens* presentes apenas na extração do ChatGPT ou Gemini; e os falsos negativos são *tokens* presentes apenas na extração manual. Para tornar possível a aplicação das análises de concordância, exigiu-se também a diferenciação de variáveis. Dada a natureza dos dados farmacológicos, onde um mesmo experimento pode apresentar múltiplos protocolos de estresse ou testes comportamentais, aplicou-se para tais colunas uma lógica de rótulos múltiplos (*multi-label*), caracterizando-se por classificações não exclusivas (Tsoumakas & Katakis, 2007). Ao comparar a interseção entre os conjuntos manual e automatizados, foi possível identificar acertos parciais e evitar a penalização excessiva dos modelos. Por exemplo, para um conjunto manual contendo {"fst", "oft"} comparado ao conjunto automatizado {"fst", "tst"}, a análise de interseção identifica que metade da informação esperada foi recuperada. Nesse cenário, o modelo não é classificado como completamente incorreto, em vez disso, as métricas de precisão e sensibilidade capturam o acerto parcial ("fst") e as falhas (omissão de "oft" e inserção indevida de "tst"), resultando em

um F_1 -score que reflete o verdadeiro desempenho do modelo.

Para as variáveis de rótulo único (*single-label*), onde as classificações são mutuamente exclusivas, como é o caso das colunas marcador de neurogênese, espécie, sexo, linhagem, antidepressivo, doses do fármaco e do marcador BrdU, regimes de injeções do fármaco e do BrdU, duração do tratamento e tempo de sobrevivência, empregou-se o Índice Kappa de Cohen (κ) para medir a concordância entre a extração manual e as automatizadas em estudos pareados pelo algoritmo *best-fit*. O Kappa de Cohen é considerado uma métrica robusta para testes de confiabilidade inter-observadores (Cohen, 1960), oferecendo um rigor adicional ao penalizar a concordância ao acaso.

Kappa de Cohen (κ):

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)}$$

Em que:

- κ : Índice Kappa de Cohen;
- p_o : concordância entre os observadores;
- p_e : concordância esperada ao acaso;
- $1 - p_e$: concordância máxima possível que não é fruto do acaso.

A força da concordância obtida foi classificada com base nos intervalos propostos por Landis e Koch (1977), conforme apresentado no **Quadro 2**.

Quadro 2. Escala de interpretação do Índice Kappa (κ).

Valor de κ	Força da Concordância
< 0,00	Ausente
0,00 – 0,20	Insignificante
0,21 – 0,40	Discreta
0,41 – 0,60	Moderada
0,61 – 0,80	Substancial
0,81 – 1,00	Quase perfeita

Adaptado de Landis e Koch (1977).

Para visualizar e diagnosticar os padrões de erro cometidos pelos modelos ChatGPT e Gemini, foram geradas matrizes de confusão (**Figura 7**), técnica de tabulação cruzada que permite o mapeamento visual das discordâncias e a identificação de vieses sistemáticos na classificação (Grandini et al., 2020). A matriz confronta as categorias do padrão-ouro, dispostas nas linhas, com as previsões efetuadas pela IA, dispostas nas colunas. A interpretação da matriz se baseia na análise de seus eixos, em que a diagonal principal concentra as instâncias de concordância perfeita, ou seja, os verdadeiros positivos. Dessa forma, quanto maior a densidade de valores observada ao longo desta diagonal, maior é a precisão demonstrada pelo modelo para a variável em questão. Inversamente, as células dispostas fora da diagonal revelam a natureza e a distribuição dos erros cometidos; ao observar as áreas de dispersão desses valores, torna-se possível identificar se o modelo apresenta dificuldades em distinguir categorias semanticamente próximas — como a confusão entre linhagens de roedores com características similares ou classes de fármacos análogos — expondo, assim, vieses sistemáticos. Dessa forma, a matriz de confusão atua como uma ferramenta qualitativa, permitindo compreender não apenas quanto o modelo errou, mas onde ele cometeu o erro.

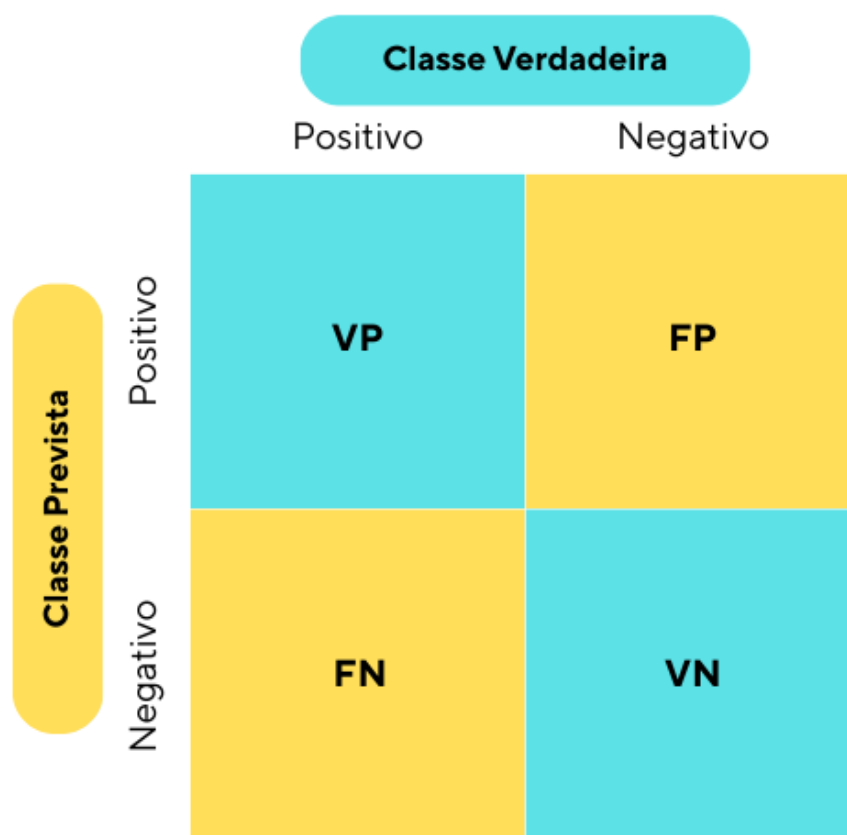


Figura 7. Estrutura de uma matriz de confusão. Os verdadeiros positivos (VP) indicam acertos na presença de um valor (e.g., identificação correta no manual e na IA), enquanto os verdadeiros negativos (VN) indicam a concordância na ausência de um valor. Já os falsos positivos (FP) representam inclusões indevidas pela IA e os falsos negativos (FN) representam omissões de dados. *Adaptado de Scikit-Learn Developers (2025).*

Finalmente, a integridade estrutural do protocolo foi avaliada pela divergência de contagem, comparando se o número total de estudos extraídos por artigo era equivalente entre os métodos manual e automatizado, o que gerou gráficos comparativos *pré* e *pós-alinhamento*. O *status* de alinhamento serviu para validar o pareamento de entidades, quantificando quantos estudos foram correspondidos (pareados) e quantos remanesceram como omissões ou alucinações. Adicionalmente, conduziu-se uma análise de erros induzidos pelo desalinhamento para classificar e ranquear a incidência de falsos positivos e falsos negativos estruturais.

5 RESULTADOS

5.1 Estruturação e padronização do protocolo de extração automatizada

O processo de estruturação e padronização do protocolo necessitou de aproximadamente quatro meses para ser concluído, considerando-se o tempo para levantamento de informações a respeito dos modelos de IA e para a engenharia de *prompt*, incluindo o refinamento por meio de extrações repetitivas até a versão consolidada do *prompt* para extração definitiva. Com base na aplicação dos critérios de seleção dos LLMs previamente estabelecidos, os modelos GPT-5 (gratuito) e Gemini 2.5 Pro (licença educacional) foram considerados os mais adequados aos propósitos deste estudo devido à acessibilidade e à capacidade de leitura massiva de artigos, respectivamente (**Tabela 1**).

O *prompt* continha comandos para a extração das variáveis descritas anteriormente na seção 4.3.1. A versão inicial do *prompt* teve uma taxa de resposta muito menor do que a versão final, com poucos ou nenhum estudo sendo retornado por artigo devido à falta de instrução específica para que o modelo extraísse a informação de interesse (**Quadro 3**). Nessa versão inicial, os comandos eram em texto corrido e sem instruções diretivas, exigindo “interpretação” do modelo de IA. Para obter a estrutura final do *prompt*, foram aplicadas restrições de alucinação implementadas através de diretrizes proibitivas explícitas e determinando, por exemplo, o uso da sigla “NR” (*not reported*) para dados não explícitos no texto, visando restringir a tendência de preenchimento probabilístico e reduzir alucinações pelo modelo. Além disso, foram inseridas instruções para priorizar unidades de medida padrão (e.g., mg/kg), de modo a facilitar a tabulação dos dados, bem como para exemplificar os padrões farmacológicos de dosagem. Houve ainda instruções de filtragem para forçar o modelo a validar apenas fármacos encontrados na lista especificada. O *prompt* alcançou sua versão final por saturação da extração. Os seguintes critérios foram aplicados para encerrar o refinamento e definir o *prompt* final: 1) interrupção do preenchimento de “NR” nas colunas que continham dados a

serem extraídos; 2) aumento no número de estudos identificados por artigo (de 1 a 2 estudos para >4 estudos); 3) estrutura praticamente igual entre as iterações, sem adições significativas de novas informações extraídas por meio de ajustes no *prompt*; 4) agrupamento de múltiplos estudos, separados por vírgulas na mesma célula, indicando maior captação de dados mesmo que sem separação de estudos por linha. O *prompt* consolidado utilizado para a extração subsequente é apresentado no **Quadro 4** e pode ser acessado no [OSF](#) (em “Files/Extração/Dissertação_Hamze_2026/Scripts”).

Tabela 1. Comparativo entre os Grandes Modelos de Linguagem (LLMs)

Modelo	Desenvolvedor	Janela de Contexto*	Assinatura (mensal)	Custo da API**	Vantagens	Limitações
GPT-5	OpenAI	128.000	Gratuito: limitado.			
			Plus (\$20): acesso geral.	Entrada: \$5,00	Alta consistência na estruturação de dados e robustez semântica para textos complexos.	Janela de contexto menor, custo alto para grandes bibliotecas.
			Pro (\$200): maior limite de uso.	Saída: \$15,00		
Gemini 2.5 Pro	Google	2.000.000+	Gratuito: limitado.			
			Advanced (\$25): acesso geral.	Entrada: \$1,25	Maior janela de contexto, integração nativa multimodal.	Limite de requisições por minuto (RPM) no plano gratuito.
			Ultra (\$249,99): capacidade máxima de uso.	Saída: \$2,50		
Llama 3	Meta	128.000	Gratuito: licença de uso livre.	Sem custo de API.	Privacidade de dados (uso local), ausência de custos recorrentes por <i>tokens</i> .	Requer infraestrutura de <i>hardware</i> (requisito mínimo: GPU 24GB + VRAM).
			Não possui plano de assinatura.	Custo de <i>hardware</i> .		
Claude 4 Opus	Anthropic	200.000	Gratuito: muito limitado.			
			Pro (\$20): acesso moderado.	Entrada: \$15,00	Excelência em raciocínio lógico e nuances linguísticas, menor taxa de alucinação.	Custo muito elevado para triagem de artigos, poucas mensagens por dia no plano gratuito.
			Max (\$100): acesso total e alto desempenho.	Saída: \$75,00		

(*) Janela de Contexto: define a quantidade máxima de informação (*tokens*) que o modelo consegue processar em uma única interação. (**) Valores expressos em dólares americanos (USD) para cada 1 milhão (1M) de *tokens* processados. Entrada (*input*): refere-se ao custo de leitura de informação enviada; saída (*output*): refere-se ao custo da resposta gerada pela IA. VRAM: memória de vídeo necessária na GPU para rodar o modelo localmente. Descrições de preços, planos e especificações técnicas em vigência até dezembro de 2025, sujeitas a alterações pelos desenvolvedores. *Elaborado pela autora (2025).*

Quadro 3. Estrutura inicial do *prompt* para extração de dados em revisão sistemática de farmacologia com base na ferramenta PICO

Extract the following information from the attached PDF, keeping your answers straight and only within what is requested:

- 1. Author: Enter only the last name of the first author listed and use the term "et al" to represent when there are more than two authors. If there are exactly two authors, provide the last name of both authors.*
- 2. Year of publication:*
- 3. Study: Number each study conducted in the article, and only consider that one study is different from another when the species, lineage, antidepressant, dose, duration of treatment, and other differences are detected. If more than one study was conducted in the same article, from here you must extract the information separately for each study.*
- 4. Species: State whether the animal in the study is rat or mice.*
- 5. Sex: State whether the animal in the study is male or female.*
- 6. Strain: State the strain of the animal in the study. Some examples of strains for rats are Wistar and Sprague-Dawley, and for mice are Swiss, C57BL/6, and BALB/c.*
- 7. Antidepressant: State which antidepressant was used in the study, which must be one of the following: agomelatine, amitriptyline, bupropion, citalopram, clomipramine, desvenlafaxine, duloxetine, escitalopram, fluoxetine, fluvoxamine, levomilnacipran, milnacipran, mirtazapine, nefazodone, paroxetine, reboxetine, sertraline, trazodone, venlafaxine, vilazodone, vortioxetine.*
- 8. Antidepressant dose: Enter dose in mg/kg of the antidepressant used. If the information is not in mg/kg, report the dose as stated in the study.*
- 9. Duration of treatment: Report duration of treatment with antidepressant. Inform the duration according to the time measurement stated in the study.*
- 10. Stress protocol: State which stress protocol was used to generate an animal model of depression. Some examples are Chronic Mild Stress, Chronic Unpredictable Mild Stress, Maternal Separation, Social Isolation.*
- 11. Behavioral test: State whether the behavioral test used to measure effectiveness of treatment was the Tail Suspension Test and/or the Forced Swim Test.*
- 12. Behavioral data: Enter measurements resulting from behavioral tests, that is, immobility time in seconds (s).*

Elaborado pela autora (2025).

Quadro 4. Estrutura final do *prompt* para extração de dados em revisão sistemática de farmacologia com base na ferramenta PICO

Componente	Comando de Instrução (<i>Prompt</i>)
Contexto	CONTEXT: You are performing data extraction in scientific literature for a systematic review and meta-analysis on the neurogenic theory of depression. INSTRUCTIONS: Provide EXTREMELY DIRECT and CONCISE answers for spreadsheet insertion. Use only information EXPLICITLY present in the PDF. If information is not available, answer with "NR" (Not reported). IMPORTANT: Be rigorously faithful to the original text in your answers, DO NOT answer unless reported in the study.

Componente	Comando de Instrução (<i>Prompt</i>)
Bibliografia	1. AUTHOR: ➤ First author (last name only) + "et al" (if >2 authors). ➤ If exactly 2 authors: LastName1 and LastName2. ➤ Format: "Silva et al" or "Silva and Santos". 2. YEAR OF PUBLICATION: ➤ Format: YYYY (e.g., 2023).
Estudos	3. STUDY IDENTIFICATION: ➤ Number each independent experimental study INSIDE the article (as study 1, 2, 3...). ➤ Consider studies in the same article different when there are variations on experimental design: neurogenesis marker, species, sex, strain, antidepressant, antidepressant dose, treatment duration, injections, BrdU parameters. ➤ If an article contains multiple studies, create ONE LINE for EACH study.
População	4. NEUROGENESIS MARKER: ➤ Examples: DCX, BrdU, KI-67, NeuN. ➤ Answer: Exact marker name. 5. SPECIES: ➤ Answer: "Rat" or "Mice" or "Mouse". 6. SEX: ➤ Answer: "Male" or "Female". 8. STRAIN: ➤ Rat examples: Wistar, Sprague-Dawley. ➤ Mice examples: Swiss, C57BL/6, BALB/c. ➤ Answer: Exact strain name.
Intervenção	9. ANTIDEPRESSANT: ➤ ONLY if one of these: agomelatine, amitriptyline, bupropion, citalopram, clomipramine, desvenlafaxine, duloxetine, escitalopram, fluoxetine, fluvoxamine, levomilnacipran, imipramine, milnacipran, mirtazapine, nefazodone, paroxetine, reboxetine, sertraline, trazodone, venlafaxine, vilazodone, vortioxetine. ➤ If not on the list, answer with "NR". 10. ANTIDEPRESSANT DOSE: ➤ Priority: mg/kg. ➤ If other unit: report exactly as in the study. 11. TREATMENT DURATION: ➤ Use exact unit from study (e.g., "14 days"). 12. INJECTIONS (per day): ➤ Answer number of injections exactly as reported.
BrdU	13. BRDU INJECTIONS: ➤ Answer number of injections exactly as reported. 14. BRDU DOSE: ➤ Priority: mg/kg.

Componente	Comando de Instrução (<i>Prompt</i>)
	15. SURVIVAL TIME FOR BRDU: ➤ Survival time AFTER last BrdU injection. ➤ Format: "14 days", "24 hours".
Desfecho	16. STRESS PROTOCOL: ➤ Examples: Chronic Mild Stress, Maternal Separation. ➤ Answer: Exact stress protocol name. 17. BEHAVIORAL TEST: ➤ Examples: Forced Swim Test, Open Field Test. ➤ Answer: Exact behavioral test name.
Formato de Saída	RESPONSE FORMAT: ➤ Present data in simple tabular format. ➤ Column order: Author, Year, Study, Neurogenesis marker, Species, Sex, Strain, Antidepressant (ATD), ATD Dose, ADT treatment duration, ATD injections, BrdU injections, BrdU dose, BrdU survival time, Stress protocol, Behavioral test.

Elaborado pela autora (2025).

5.2 Extração de dados quantitativos e qualitativos GPT ou Gemini vs. Manual

O fluxo de trabalho no ChatGPT e no Gemini diferiram devido às limitações impostas pelas regras de uso dos planos, causando também um impacto na estrutura final dos dados brutos. As três planilhas com os dados brutos referentes às extrações Manual, ChatGPT e Gemini estão disponíveis no [OSF](#) (em “Files/Extração/Dissertação_Hamze_2026/Dados”).

Em relação ao tempo gasto no processo de extração (**Tabela 2**), a extração manual demandou um trabalho estimado em 1 hora por artigo, totalizando aproximadamente 26 horas de trabalho ativo para a conclusão da amostra. No cenário realista, essa carga horária foi diluída ao longo de 12 semanas (cerca de 2-3 artigos por semana) devido à necessidade de pausas da primeira avaliadora (FCH). Ao final da extração e após a conciliação, uma única planilha foi estruturada contendo 74 linhas (estudos) para a síntese de dados, denominada como “Planilha Manual”. A extração automatizada usando o ChatGPT reduziu drasticamente o tempo de execução da tarefa, desempenhando a mesma função em um intervalo médio de 10 segundos. Entretanto, apesar da velocidade do modelo em extrair dados, o tempo total decorrido para concluir a tarefa foi de 3 dias, devido ao fracionamento da amostra em três ciclos de extração. Os ciclos resultaram em três planilhas parciais, as quais foram unificadas em um único arquivo Excel (.xlsx) para a síntese de dados, contendo 35 linhas e sendo nomeada como “Planilha GPT”. Semelhante ao ChatGPT, a extração de dados usando o Gemini também gerou planilhas de dados em um tempo médio de 10 segundos, mas sem a necessidade de fracionamento de

artigos. Neste caso, o Gemini 2.5 Pro permitiu maior eficiência de fluxo e permitiu o processamento da amostra completa em um ciclo único. A extração pelo Gemini resultou em uma planilha contendo 58 linhas, nomeada como “Planilha Gemini”.

Tabela 2. Tempo gasto na extração de dados da amostra de artigos

	Tempo por artigo (aproximado)	Tempo total (n=26 artigos)	Tempo para a conclusão (dados tabulados)
Manual	3.600s (1 h)	93.600s (26 h)	12 semanas (2-3 artigos por semana)
ChatGPT (gratuito)	0,2-0,6s	30s (3 ciclos)	3 dias (1 ciclo de ~10s por dia)
Gemini Pro 2.5	0,2-0,6s	10s (1 ciclo)	10s

Elaborado pelo autora (2025).

A análise dos arquivos brutos revelou uma diferença na estruturação dos dados em comparação com a Planilha Manual (**Tabela 3**), observando-se que ambos os modelos apresentaram uma tendência a aglutinar dados por agrupar múltiplos estudos na mesma célula. Portanto, o tratamento dos dados foi necessário para obter uma estrutura analítica padronizada (*tidy data*) (Wickham, 2014). A aplicação da função *explode* resultou na expansão das bases de dados, aumentando o *dataframe* do ChatGPT de 35 linhas para 55 linhas, do Manual de 74 linhas para 96 linhas e do Gemini de 58 linhas para 78 linhas. Embora o tratamento de dados tenha aumentado o volume de estudos, ainda se observou que em relação à planilha manual havia 41 e 18 estudos a menos para o ChatGPT e para Gemini, respectivamente, sugerindo que a divergência não decorre apenas da formatação aglutinada, mas também da omissão de estudos. Os dados limpos e tratados foram armazenados em novos *dataframes* padronizados, constituindo a base para o alinhamento e para as análises estatísticas subsequentes.

Tabela 3. Estrutura da tabulação dos dados extraídos pré e pós-tratamento

	Linhas (estudos) pré-tratamento	Linhas (estudos) pós-tratamento	Diferença entre o número de estudos identificados
Manual	74	96	22
ChatGPT (gratuito)	35	55	20
Gemini Pro 2.5	58	78	20

Elaborado pela autora (2025).

5.2.1 Extração de variáveis para a população

A **Tabela 4** apresenta uma comparação entre as variáveis espécie, sexo e linhagem extraídos pela forma tradicional (manual) vs. ChatGPT e Gemini. Com o objetivo de facilitar a

clareza das informações, as tabelas foram separadas por componentes do PICO e foram usados os dados tratados de modo agrupado para melhor visualização e fluidez de leitura. Os IDs dos artigos foram mantidos como aparecem no banco de dados original para fins de rastreabilidade.

Tabela 4. Variáveis para a população (P) das extrações Manual, GPT e Gemini

Origem	ID	Estudo	Espécie	Sexo	Linhagem
Manual	1	1	mice	male	swiss
GPT	1	1	mice	male	swiss
GPT	1	2	mice	male	swiss
GPT	1	3	mice	male	swiss
Gemini	1	1	mice	male	swiss
Manual	12	1	rat	male	wistar
GPT	12	1	rat	male	wistar
GPT	12	2	rat	male	wistar
Gemini	12	1	rat	male	wistar
Gemini	12	2	rat	male	wistar
Manual	15	1	mice	male	swiss
Manual	15	2	mice	male	swiss
Manual	15	3	mice	male	swiss
Manual	15	4	mice	male	swiss
Manual	15	5	mice	male	swiss
Manual	15	6	mice	male	swiss
Manual	15	7	mice	male	swiss
Manual	15	8	mice	male	swiss
GPT	15	1	mice	male	swiss
GPT	15	2	mice	male	swiss
GPT	15	3	mice	male	swiss
GPT	15	4	mice	male	swiss
GPT	15	5	mice	male	swiss
GPT	15	6	mice	male	swiss
GPT	15	7	mice	male	swiss
GPT	15	8	mice	male	swiss
GPT	15	9	mice	male	swiss
GPT	15	10	mice	male	swiss
Gemini	15	1	mice	male	swiss
Gemini	15	2	mice	male	swiss
Gemini	15	3	mice	male	swiss
Gemini	15	4	mice	male	swiss
Manual	16	1	mice	male	c57bl/6
GPT	16	1	mice	male	nr
Gemini	16	1	mice	male	c57bl/6
Manual	18	1	mice	male	balb/c
Manual	18	2	mice	male	balb/c
GPT	18	1	mice	nr	nr
Gemini	18	1	mice	male	balb/c
Gemini	18	2	mice	male	balb/c
Manual	19	1	mice	male	cd-1
Manual	19	2	mice	male	cd-1
GPT	19	1	mice	male	cd-1
GPT	19	2	mice	male	cd-1
GPT	19	3	mice	male	cd-1
Gemini	19	1	mice	male	cd-1
Gemini	19	2	mice	male	hr line
Gemini	19	3	mice	male	cd-1
Gemini	19	4	mice	male	hr line
Gemini	19	5	mice	male	cd-1
Gemini	19	6	mice	male	ir line
Gemini	19	7	mice	male	cd-1
Gemini	19	8	mice	male	ir line
Gemini	19	9	mice	male	cd-1
Gemini	19	10	mice	male	ir line
Gemini	19	11	mice	male	cd-1

Gemini	19	12	mice	male	ir line
Manual	21	1	mice	male	balb/c
GPT	21	1	mice	nr	balb/c
Gemini	21	1	mice	male	balb/c
Manual	25	1	mice	male	icr
GPT	25	1	mice	male	icr
GPT	25	2	mice	male	icr
Gemini	25	1	mice	male	icr
Gemini	25	2	mice	male	icr
Manual	28	1	rat	male	wistar
GPT	28	1	rat	male	wistar
Gemini	28	1	rat	male	wistar
Manual	30	1	mice	male	icr
GPT	30	1	mice	male	icr
GPT	30	2	mice	male	icr
Gemini	30	1	mice	male	icr
Gemini	30	2	mice	male	icr
Gemini	30	3	mice	male	icr
Manual	31	1	mice	male	c57bl/6
GPT	31	1	mice	male	c57bl/6
Gemini	31	1	mice	male	c57bl/6
Manual	32	1	mice	male	camkiv
Manual	32	2	mice	male	camkiv
GPT	32	1	mice	nr	camkiv
Gemini	32	1	mice	male	129s1
Gemini	32	2	mice	male	camkiv
Gemini	32	3	mice	male	129s1
Gemini	32	4	mice	male	camkiv
Manual	39	1	mice	male	std-ddy
Manual	39	2	mice	male	std-ddy
GPT	39	1	mice	nr	nr
GPT	39	2	mice	nr	nr
Gemini	39	1	mice	male	std-ddy
Gemini	39	2	mice	male	std-ddy
Gemini	39	3	mice	male	std-ddy
Gemini	39	4	mice	male	std-ddy
Gemini	39	5	mice	male	std-ddy
Gemini	39	6	mice	male	std-ddy
Manual	40	1	mice	male	balb/c
GPT	40	1	mice	nr	nr
Gemini	40	1	mice	male	balb/c
Manual	43	1	mice	male	kun-ming
GPT	43	1	mice	male	nr
Gemini	43	1	mice	male	kun-ming
Gemini	43	2	mice	male	kun-ming
Manual	45	1	rat	male	sprague-dawley
GPT	45	1	rat	male	sprague-dawley
GPT	45	2	rat	male	sprague-dawley
Gemini	45	1	rat	male	sprague-dawley
Gemini	45	2	rat	male	sprague-dawley
Manual	48	1	rat	male	sprague-dawley
Manual	48	2	rat	male	sprague-dawley
GPT	48	1	rat	male	sprague-dawley
GPT	48	2	rat	male	sprague-dawley
GPT	48	3	rat	male	sprague-dawley
Gemini	48	1	rat	male	sprague-dawley
Gemini	48	2	rat	male	sprague-dawley
Gemini	48	3	rat	male	sprague-dawley
Gemini	48	4	rat	male	sprague-dawley
Manual	52	1	mice	both	euploid
Manual	52	2	mice	both	euploid
Manual	52	3	mice	both	ts65dn
Manual	52	4	mice	both	ts65dn
Manual	52	5	mice	both	euploid
Manual	52	6	mice	both	euploid
Manual	52	7	mice	both	ts65dn
Manual	52	8	mice	both	ts65dn

Manual	52	9	mice	both	euploid
Manual	52	10	mice	both	euploid
Manual	52	11	mice	both	ts65dn
Manual	52	12	mice	both	ts65dn
Manual	52	13	mice	both	euploid
Manual	52	14	mice	both	euploid
Manual	52	15	mice	both	ts65dn
Manual	52	16	mice	both	ts65dn
Manual	52	17	mice	both	euploid
Manual	52	18	mice	both	euploid
Manual	52	19	mice	both	ts65dn
Manual	52	20	mice	both	ts65dn
Manual	52	21	mice	both	euploid
Manual	52	22	mice	both	euploid
Manual	52	23	mice	both	ts65dn
Manual	52	24	mice	both	ts65dn
Manual	52	25	mice	both	euploid
Manual	52	26	mice	both	euploid
Manual	52	27	mice	both	ts65dn
Manual	52	28	mice	both	ts65dn
Manual	52	29	mice	both	euploid
Manual	52	30	mice	both	euploid
Manual	52	31	mice	both	ts65dn
Manual	52	32	mice	both	ts65dn
Manual	52	33	mice	both	euploid
Manual	52	34	mice	both	euploid
Manual	52	35	mice	both	ts65dn
Manual	52	36	mice	both	ts65dn
Manual	52	37	mice	both	euploid
Manual	52	38	mice	both	euploid
Manual	52	39	mice	both	ts65dn
Manual	52	40	mice	both	ts65dn
Manual	52	41	mice	both	euploid
Manual	52	42	mice	both	euploid
Manual	52	43	mice	both	ts65dn
Manual	52	44	mice	both	ts65dn
GPT	52	1	mice	nr	ts65dn
GPT	52	2	mice	nr	ts65dn
Gemini	52	1	mice	male	ts65dn
Gemini	52	2	mice	male	ts65dn
Gemini	52	3	mice	female	ts65dn
Gemini	52	4	mice	female	ts65dn
Gemini	52	5	mice	male	ts65dn
Gemini	52	6	mice	male	ts65dn
Gemini	52	7	mice	female	ts65dn
Gemini	52	8	mice	female	ts65dn
Gemini	52	9	mice	male	ts65dn
Gemini	52	10	mice	male	ts65dn
Gemini	52	11	mice	female	ts65dn
Gemini	52	12	mice	female	ts65dn
Manual	53	1	mice	male	icr
GPT	53	1	mice	nr	nr
Gemini	53	1	mice	male	c57bl/6
Manual	54	1	mice	male	dba/2j
Manual	54	2	mice	male	dba/2j
Manual	54	3	mice	male	dba/2j
Manual	54	4	mice	male	dba/2j
Manual	54	5	mice	male	dba/2j
Manual	54	6	mice	male	dba/2j
Manual	54	7	mice	male	dba/2j
Manual	54	8	mice	male	dba/2j
Manual	54	9	mice	male	dba/2j
Manual	54	10	mice	male	dba/2j
Manual	54	11	mice	male	dba/2j
Manual	54	12	mice	male	dba/2j
Manual	54	13	mice	male	129s1
Manual	54	14	mice	male	a/j

Manual	54	15	mice	male	swr/j
GPT	54	1	mice	male	129s1
GPT	54	2	mice	male	a/j
GPT	54	3	mice	male	dba/2j
GPT	54	4	mice	male	swr/j
GPT	54	5	mice	male	dba/2j
GPT	54	6	mice	male	dba/2j
Gemini	54	1	mice	male	dba/2j
Gemini	54	2	mice	male	129s1
Gemini	54	3	mice	male	a/j
Gemini	54	4	mice	male	swr/j
Gemini	54	5	mice	male	dba/2j
Gemini	54	6	mice	male	dba/2j
Manual	55	1	mice	male	kun-ming
Manual	55	2	rat	male	sprague-dawley
GPT	55	1	rat	nr	nr
Gemini	55	1	rat	male	sprague-dawley
Manual	58	1	rat	male	sprague-dawley
GPT	58	1	rat	male	sprague-dawley
Gemini	58	1	rat	male	sprague-dawley
Manual	59	1	rat	male	sprague-dawley
GPT	59	1	rat	male	nr
Gemini	59	1	rat	male	sprague-dawley
Gemini	59	2	rat	female	sprague-dawley
Manual	60	1	mice	male	c57bl/6
GPT	60	1	mice	nr	c57bl/6
Gemini	60	1	mice	male	c57bl/6
Gemini	60	2	mice	male	c57bl/6
Manual	61	1	mice	male	balb/c
GPT	61	1	mice	nr	balb/c
GPT	61	2	mice	nr	balb/c
Gemini	61	1	mice	male	balb/c
Gemini	61	2	mice	male	balb/c
Gemini	61	3	mice	male	balb/c
Manual	62	1	mice	male	c57bl/6
GPT	62	1	mice	male	c57bl/6
GPT	62	2	mice	male	c57bl/6
GPT	62	3	mice	male	c57bl/6
Gemini	62	1	mice	male	c57bl/6

NR: não reportado. *Elaborado pela autora (2025).*

5.2.2 Extração de variáveis para a intervenção

A **Tabela 5** apresenta uma comparação entre as variáveis **antidepressivo**, **dose** e **número de injeções** e **duração do tratamento**, além dos dados de **proliferação** e **sobrevivência** (eutanásia) para o marcador BrdU, extraídos pela forma tradicional (manual) vs. ChatGPT e Gemini.

Tabela 5. Variáveis para a intervenção (I) das extrações Manual, GPT e Gemini

Origem	ID	Estudo	Antidepressivo				Marcador de Neurogênese BrdU			
			Tipo	Dose	Duração	Injeção	Tipo	Injeção	Dose	Eutanásia
Manual	1	1	flx	20 mg/kg	14 d	1	ki-67	nr	nr	nr
GPT	1	1	flx	nr	28 d	nr	ki-67	nr	nr	nr
GPT	1	2	nr	nr	28 d	nr	ki-67	nr	nr	nr
GPT	1	3	nr	nr	28 d	nr	bdnf	nr	nr	nr
Gemini	1	1	flx	20 mg/kg	14 d	1	ki-67	nr	nr	nr
Manual	12	1	flx	5 mg/kg	15 d	1	brdu	3	18 mg/ml	1 d
GPT	12	1	flx	5 mg/kg	15 d	1	brdu	3	75 mg/kg	1 d
GPT	12	2	flx	5 mg/kg	15 d	1	dcx	3	75 mg/kg	1 d
Gemini	12	1	flx	5 mg/kg	15 d	1	brdu	3	75 mg/kg	1 d

Gemini	12	2	flx	5 mg/kg	15 d	1	dcx	nr	nr	nr
Manual	15	1	imi	30 mg/kg	2 h	1	ki-67	nr	nr	nr
Manual	15	2	imi	30 mg/kg	1 d	1	ki-67	nr	nr	nr
Manual	15	3	imi	0.06 nmol	2 h	1	ki-67	nr	nr	nr
Manual	15	4	imi	0.06 nmol	1 d	1	ki-67	nr	nr	nr
Manual	15	5	imi	0.2 nmol	2 h	1	ki-67	nr	nr	nr
Manual	15	6	imi	0.2 nmol	1 d	1	ki-67	nr	nr	nr
Manual	15	7	imi	0.2 nmol	2 h	1	dcx	nr	nr	nr
Manual	15	8	imi	0.06 nmol	1 d	1	dcx	nr	nr	nr
GPT	15	1	nr	nr	nr	nr	ki-67	nr	nr	nr
GPT	15	2	nr	nr	nr	nr	dcx	nr	nr	nr
GPT	15	3	imi	30 mg/kg	1 d	1	ki-67	nr	nr	nr
GPT	15	4	imi	30 mg/kg	1 d	1	dcx	nr	nr	nr
GPT	15	5	imi	0.2 nmol	1 d	1	ki-67	nr	nr	nr
GPT	15	6	imi	0.2 nmol	1 d	1	dcx	nr	nr	nr
GPT	15	7	imi	0.06 nmol	1 d	1	ki-67	nr	nr	nr
GPT	15	8	imi	0.2 nmol	1 d	1	ki-67	nr	nr	nr
GPT	15	9	imi	0.06 nmol	1 d	1	dcx	nr	nr	nr
GPT	15	10	imi	0.2 nmol	1 d	1	dcx	nr	nr	nr
Gemini	15	1	imi	30 mg/kg	acute	1	ki-67	nr	nr	nr
Gemini	15	2	imi	0.06 nmol	acute	1	ki-67	nr	nr	nr
Gemini	15	3	imi	0.2 nmol	acute	1	ki-67	nr	nr	nr
Gemini	15	4	imi	0.2 nmol	acute	1	dcx	nr	nr	nr
Manual	16	1	esc	10 mg/kg	21 d	1	dcx	nr	nr	nr
GPT	16	1	esc	10 mg/kg	21 d	1	dcx	nr	nr	nr
Gemini	16	1	esc	10 mg/kg	21 d	1	dcx	nr	nr	nr
Manual	18	1	flx	10 mg/kg	27 d	1	brdu	12	300 mg/kg	1 d
Manual	18	2	imi	20 mg/kg	27 d	1	brdu	12	300 mg/kg	1 d
GPT	18	1	nr	nr	nr	nr	brdu	4	75 mg/kg	1 d
Gemini	18	1	flx	10 mg/kg	28 d	1	brdu	4	75 mg/kg	1 d
Gemini	18	2	imi	20 mg/kg	28 d	1	brdu	4	75 mg/kg	1 d
Manual	19	1	flx	10 mg/kg	35 d	1	dcx	nr	nr	nr
Manual	19	2	flx	10 mg/kg	35 d	1	ki-67	nr	nr	nr
GPT	19	1	flx	10 mg/kg	35 d	1	nr	nr	nr	nr
GPT	19	2	flx	10 mg/kg	33 d	1	ki-67	nr	nr	nr
GPT	19	3	flx	10 mg/kg	33 d	1	dcx	nr	nr	nr
Gemini	19	1	flx	10 mg/kg	33 d	1	ki-67	nr	nr	nr
Gemini	19	2	flx	10 mg/kg	33 d	1	ki-67	nr	nr	nr
Gemini	19	3	flx	10 mg/kg	33 d	1	dcx	nr	nr	nr
Gemini	19	4	flx	10 mg/kg	33 d	1	dcx	nr	nr	nr
Gemini	19	5	flx	10 mg/kg	33 d	1	ki-67	nr	nr	nr
Gemini	19	6	flx	10 mg/kg	33 d	1	ki-67	nr	nr	nr
Gemini	19	7	flx	10 mg/kg	33 d	1	dcx	nr	nr	nr
Gemini	19	8	flx	10 mg/kg	33 d	1	dcx	nr	nr	nr
Gemini	19	9	flx	10 mg/kg	33 d	1	ki-67	nr	nr	nr
Gemini	19	10	flx	10 mg/kg	33 d	1	ki-67	nr	nr	nr
Gemini	19	11	flx	10 mg/kg	33 d	1	dcx	nr	nr	nr
Gemini	19	12	flx	10 mg/kg	33 d	1	dcx	nr	nr	nr
Manual	21	1	flx	15 mg/kg	42 d	1	dcx	nr	nr	nr
GPT	21	1	flx	15 mg/kg	nr	nr	dcx	nr	nr	nr
Gemini	21	1	flx	15 mg/kg	35 d	nr	dcx	nr	nr	nr
Manual	25	1	flx	10 mg/kg	21 d	1	dcx	nr	nr	nr
GPT	25	1	flx	10 mg/kg	21 d	1	ki-67	nr	nr	nr
GPT	25	2	flx	10 mg/kg	21 d	1	dcx	nr	nr	nr
Gemini	25	1	flx	10 mg/kg	21 d	1	ki-67	nr	nr	nr
Gemini	25	2	flx	10 mg/kg	21 d	1	dcx	nr	nr	nr
Manual	28	1	flx	18 mg/kg	24 d	1	brdu	4	50 mg/kg	4 d
GPT	28	1	flx	18 mg/kg	24 d	1	brdu	4	50 mg/kg	nr
Gemini	28	1	flx	18 mg/kg	24 d	1	brdu	4	50 mg/kg	3 d
Manual	30	1	flx	15 mg/kg	21 d	1	brdu	3	150 mg/kg	23 d
GPT	30	1	flx	15 mg/kg	nr	nr	dcx	3	150 mg/kg	21 d
GPT	30	2	flx	15 mg/kg	nr	nr	brdu	3	150 mg/kg	21 d
Gemini	30	1	flx	15 mg/kg	21 d	1	dcx	nr	nr	nr
Gemini	30	2	flx	15 mg/kg	21 d	1	brdu	3	150 mg/kg	20 d
Gemini	30	3	flx	15 mg/kg	21 d	1	neun	3	150 mg/kg	20 d
Manual	31	1	esc	10 mg/kg	21 d	1	dcx	nr	nr	nr
GPT	31	1	nr	nr	nr	nr	dcx	nr	nr	nr

Gemini	31	1	esc	10 mg/kg	21 d	1	dcx	nr	nr	nr
Manual	32	1	prx	1 mg/kg	14 d	1	brdu	nr	nr	nr
Manual	32	2	flv	2.5 mg/kg	14 d	1	brdu	nr	nr	nr
GPT	32	1	flv	nr	14 d	nr	brdu	nr	nr	nr
Gemini	32	1	flv	2.5 mg/kg	14 d	nr	brdu	5	nr	nr
Gemini	32	2	flv	2.5 mg/kg	14 d	nr	brdu	5	nr	nr
Gemini	32	3	prx	1.0 mg/kg	14 d	nr	brdu	5	nr	nr
Gemini	32	4	prx	1.0 mg/kg	14 d	nr	brdu	5	nr	nr
Manual	39	1	flv	30 mg/kg	14 d	1	brdu	28	50 mg/kg	12 h
Manual	39	2	imi	30 mg/kg	14 d	1	brdu	28	50 mg/kg	12 h
GPT	39	1	nr	nr	nr	nr	brdu	2	50 mg/kg	12 h
GPT	39	2	nr	nr	nr	nr	nestin	2	50 mg/kg	12 h
Gemini	39	1	imi	30 mg/kg	14 d	1	brdu	2	50 mg/kg	12 h
Gemini	39	2	flv	30 mg/kg	14 d	1	brdu	2	50 mg/kg	12 h
Gemini	39	3	flv	30 mg/kg	21 d	1	brdu	10	50 mg/kg	14 d
Gemini	39	4	flv	30 mg/kg	21 d	1	neun	10	50 mg/kg	14 d
Gemini	39	5	flv	30 mg/kg	28 d	1	brdu	10	50 mg/kg	21 d
Gemini	39	6	flv	30 mg/kg	28 d	1	neun	10	50 mg/kg	21 d
Manual	40	1	esc	10 mg/kg	7 d	1	dcx	nr	nr	nr
GPT	40	1	nr	nr	nr	nr	dcx	nr	nr	nr
Gemini	40	1	esc	10 mg/kg	7 d	1	dcx	nr	nr	nr
Manual	43	1	imi	10 mg/kg	24 d	1	brdu	4	100 mg/kg	1 d
GPT	43	1	nr	nr	nr	nr	brdu	4	100 mg/kg	1 d
Gemini	43	1	imi	10 mg/kg	24 d	1	brdu	4	100 mg/kg	1 d
Gemini	43	2	nr	nr	nr	nr	brdu	4	100 mg/kg	28 d
Manual	45	1	cit	10 mg/kg	28 d	1	brdu	4	50 mg/kg	13 d
GPT	45	1	cit	nr	nr	nr	brdu	nr	nr	nr
GPT	45	2	cit	nr	nr	nr	neun	nr	nr	nr
Gemini	45	1	cit	10 mg/kg	28 d	1	brdu	4	50 mg/kg	13 d
Gemini	45	2	cit	10 mg/kg	28 d	1	neun	4	50 mg/kg	13 d
Manual	48	1	imi	20 mg/kg	35 d	1	brdu	2	100 mg/kg	1 d
Manual	48	2	imi	20 mg/kg	35 d	1	dcx	nr	nr	nr
GPT	48	1	imi	20 mg/kg	35 d	1	brdu	nr	nr	nr
GPT	48	2	imi	20 mg/kg	35 d	1	neun	nr	nr	nr
GPT	48	3	imi	20 mg/kg	35 d	1	dcx	nr	nr	nr
Gemini	48	1	imi	20 mg/kg	35 d	1	bdnf	nr	nr	nr
Gemini	48	2	imi	20 mg/kg	35 d	1	dcx	nr	nr	nr
Gemini	48	3	imi	20 mg/kg	35 d	1	brdu	2	100 mg/kg	1 d
Gemini	48	4	imi	20 mg/kg	35 d	1	neun	2	100 mg/kg	1 d
Manual	52	1	flx	5 mg/kg	12 d	1	brdu	1	150 ug/g	2 h
Manual	52	2	flx	10 mg/kg	12 d	1	brdu	1	150 ug/g	2 h
Manual	52	3	flx	5 mg/kg	12 d	1	brdu	1	150 ug/g	2 h
Manual	52	4	flx	10 mg/kg	12 d	1	brdu	1	150 ug/g	2 h
Manual	52	5	flx	5 mg/kg	12 d	1	brdu	1	150 ug/g	2 h
Manual	52	6	flx	10 mg/kg	12 d	1	brdu	1	150 ug/g	2 h
Manual	52	7	flx	5 mg/kg	12 d	1	brdu	1	150 ug/g	2 h
Manual	52	8	flx	10 mg/kg	12 d	1	brdu	1	150 ug/g	2 h
Manual	52	9	flx	5 mg/kg	12 d	1	brdu	1	150 ug/g	30 d
Manual	52	10	flx	10 mg/kg	12 d	1	brdu	1	150 ug/g	30 d
Manual	52	11	flx	5 mg/kg	12 d	1	brdu	1	150 ug/g	30 d
Manual	52	12	flx	10 mg/kg	12 d	1	brdu	1	150 ug/g	30 d
Manual	52	13	flx	5 mg/kg	12 d	1	brdu	1	150 ug/g	30 d
Manual	52	14	flx	10 mg/kg	12 d	1	brdu	1	150 ug/g	30 d
Manual	52	15	flx	5 mg/kg	12 d	1	brdu	1	150 ug/g	30 d
Manual	52	16	flx	10 mg/kg	12 d	1	brdu	1	150 ug/g	30 d
Manual	52	17	flx	5 mg/kg	12 d	1	neun+brdu	1	150 ug/g	30 d
Manual	52	18	flx	10 mg/kg	12 d	1	neun+brdu	1	150 ug/g	30 d
Manual	52	19	flx	5 mg/kg	12 d	1	neun+brdu	1	150 ug/g	30 d
Manual	52	20	flx	10 mg/kg	12 d	1	neun+brdu	1	150 ug/g	30 d
Manual	52	21	flx	5 mg/kg	12 d	1	gfap+brdu	1	150 ug/g	30 d
Manual	52	22	flx	10 mg/kg	12 d	1	gfap+brdu	1	150 ug/g	30 d
Manual	52	23	flx	5 mg/kg	12 d	1	gfap+brdu	1	150 ug/g	30 d
Manual	52	24	flx	10 mg/kg	12 d	1	gfap+brdu	1	150 ug/g	30 d
Manual	52	25	flx	5 mg/kg	12 d	1	neither+brdu	1	150 ug/g	30 d
Manual	52	26	flx	10 mg/kg	12 d	1	neither+brdu	1	150 ug/g	30 d
Manual	52	27	flx	5 mg/kg	12 d	1	neither+brdu	1	150 ug/g	30 d
Manual	52	28	flx	10 mg/kg	12 d	1	neither+brdu	1	150 ug/g	30 d

Manual	52	29	flx	5 mg/kg	12 d	1	neun+brdu	1	150 ug/g	30 d
Manual	52	30	flx	10 mg/kg	12 d	1	neun+brdu	1	150 ug/g	30 d
Manual	52	31	flx	5 mg/kg	12 d	1	neun+brdu	1	150 ug/g	30 d
Manual	52	32	flx	10 mg/kg	12 d	1	neun+brdu	1	150 ug/g	30 d
Manual	52	33	flx	5 mg/kg	12 d	1	gfap+brdu	1	150 ug/g	30 d
Manual	52	34	flx	10 mg/kg	12 d	1	gfap+brdu	1	150 ug/g	30 d
Manual	52	35	flx	5 mg/kg	12 d	1	gfap+brdu	1	150 ug/g	30 d
Manual	52	36	flx	10 mg/kg	12 d	1	gfap+brdu	1	150 ug/g	30 d
Manual	52	37	flx	5 mg/kg	12 d	1	neither+brdu	1	150 ug/g	30 d
Manual	52	38	flx	10 mg/kg	12 d	1	neither+brdu	1	150 ug/g	30 d
Manual	52	39	flx	5 mg/kg	12 d	1	neither+brdu	1	150 ug/g	30 d
Manual	52	40	flx	10 mg/kg	12 d	1	neither+brdu	1	150 ug/g	30 d
Manual	52	41	flx	5 mg/kg	12 d	1	ki-67	1	150 ug/g	30 d
Manual	52	42	flx	10 mg/kg	12 d	1	ki-67	1	150 ug/g	30 d
Manual	52	43	flx	5 mg/kg	12 d	1	ki-67	1	150 ug/g	30 d
Manual	52	44	flx	10 mg/kg	12 d	1	ki-67	1	150 ug/g	30 d
GPT	52	1	flx	nr	12 d	nr	brdu	1	nr	28 d
GPT	52	2	flx	nr	12 d	nr	brdu	1	nr	2 h
Gemini	52	1	flx	5 mg/kg	13 d	1	brdu	1	150 mg/kg	2 h
Gemini	52	2	flx	10 mg/kg	13 d	1	brdu	1	150 mg/kg	2 h
Gemini	52	3	flx	5 mg/kg	13 d	1	brdu	1	150 mg/kg	2 h
Gemini	52	4	flx	10 mg/kg	13 d	1	brdu	1	150 mg/kg	2 h
Gemini	52	5	flx	5 mg/kg	13 d	1	brdu	1	150 mg/kg	28 d
Gemini	52	6	flx	10 mg/kg	13 d	1	brdu	1	150 mg/kg	28 d
Gemini	52	7	flx	5 mg/kg	13 d	1	brdu	1	150 mg/kg	28 d
Gemini	52	8	flx	10 mg/kg	13 d	1	brdu	1	150 mg/kg	28 d
Gemini	52	9	flx	5 mg/kg	13 d	1	ki-67	nr	nr	nr
Gemini	52	10	flx	10 mg/kg	13 d	1	ki-67	nr	nr	nr
Gemini	52	11	flx	5 mg/kg	13 d	1	ki-67	nr	nr	nr
Gemini	52	12	flx	10 mg/kg	13 d	1	ki-67	nr	nr	nr
Manual	53	1	flx	20 mg/kg	28 d	1	dcx	nr	nr	nr
GPT	53	1	nr	nr	nr	nr	dcx	nr	nr	nr
Gemini	53	1	flx	20 mg/kg	35 d	1	dcx	nr	nr	nr
Manual	54	1	flx	5 mg/kg	1 d	1	brdu	1	200 mg/kg	1 d
Manual	54	2	flx	5 mg/kg	6 d	1	brdu	1	200 mg/kg	1 d
Manual	54	3	flx	5 mg/kg	21 d	1	brdu	1	200 mg/kg	1 d
Manual	54	4	flx	10 mg/kg	1 d	1	brdu	1	200 mg/kg	1 d
Manual	54	5	flx	10 mg/kg	6 d	1	brdu	1	200 mg/kg	1 d
Manual	54	6	flx	10 mg/kg	21 d	1	brdu	1	200 mg/kg	1 d
Manual	54	7	flx	14 mg/kg	1 d	1	brdu	1	200 mg/kg	1 d
Manual	54	8	flx	14 mg/kg	6 d	1	brdu	1	200 mg/kg	1 d
Manual	54	9	flx	14 mg/kg	21 d	1	brdu	1	200 mg/kg	1 d
Manual	54	10	flx	18 mg/kg	1 d	1	brdu	1	200 mg/kg	1 d
Manual	54	11	flx	18 mg/kg	6 d	1	brdu	1	200 mg/kg	1 d
Manual	54	12	flx	18 mg/kg	21 d	1	brdu	1	200 mg/kg	1 d
Manual	54	13	flx	18 mg/kg	21 d	1	brdu	1	200 mg/kg	1 d
Manual	54	14	flx	18 mg/kg	21 d	1	brdu	1	200 mg/kg	1 d
Manual	54	15	flx	18 mg/kg	21 d	1	brdu	1	200 mg/kg	1 d
GPT	54	1	flx	18 mg/kg	21 d	water	brdu	1	200 mg/kg	1 d
GPT	54	2	flx	18 mg/kg	21 d	water	brdu	1	200 mg/kg	1 d
GPT	54	3	flx	18 mg/kg	21 d	water	brdu	1	200 mg/kg	1 d
GPT	54	4	flx	18 mg/kg	21 d	water	brdu	1	200 mg/kg	1 d
GPT	54	5	flx	18 mg/kg	21 d	water	brdu	2	200 mg/kg	3 d
GPT	54	6	flx	18 mg/kg	21 d	water	neun	2	200 mg/kg	3 d
Gemini	54	1	flx	18 mg/kg	21 d	nr	brdu	1	200 mg/kg	1 d
Gemini	54	2	flx	18 mg/kg	21 d	nr	brdu	1	200 mg/kg	1 d
Gemini	54	3	flx	18 mg/kg	21 d	nr	brdu	1	200 mg/kg	1 d
Gemini	54	4	flx	18 mg/kg	21 d	nr	brdu	1	200 mg/kg	1 d
Gemini	54	5	flx	18 mg/kg	21 d	nr	brdu	2	200 mg/kg	3 d
Gemini	54	6	flx	18 mg/kg	21 d	nr	neun	2	200 mg/kg	3 d
Manual	55	1	ami	15 mg/kg	1 d	1	brdu	nr	nr	nr
Manual	55	2	ami	10 mg/kg	14 d	nr	brdu	6	50 mg/kg	nr
GPT	55	1	nr	nr	nr	nr	brdu	6	50 mg/kg	lastinjection
Gemini	55	1	ami	10 mg/kg	14 d	1	brdu	6	50 mg/kg	nr
Manual	58	1	flx	5 mg/kg	7 d	1	brdu	8	300 mg/kg	1 d
GPT	58	1	flx	5 mg/kg	7 d	1	brdu	4	75 mg/kg	1 d
Gemini	58	1	flx	5 mg/kg	7 d	1	brdu	4	75 mg/kg	1 d

Manual	59	1	des	20 mg/kg	1 d	3	ki-67	nr	nr	nr
GPT	59	1	nr	nr	nr	nr	ki-67	nr	nr	nr
Gemini	59	1	nr	nr	1 d	3	ki-67	nr	nr	nr
Gemini	59	2	nr	nr	1 d	3	ki-67	nr	nr	nr
Manual	60	1	flx	10 mg/kg	15 d	1	ki-67	nr	nr	nr
GPT	60	1	flx	10 mg/kg	15 d	nr	ki-67	nr	nr	nr
Gemini	60	1	flx	10 mg/kg	15 d	1	ki-67	nr	nr	nr
Gemini	60	2	flx	10 mg/kg	15 d	1	neun	nr	nr	nr
Manual	61	1	flx	10 mg/kg	49 d	1	ki-67	nr	nr	nr
GPT	61	1	flx	20 mg/kg	nr	nr	ki-67	nr	nr	nr
GPT	61	2	flx	20 mg/kg	nr	nr	neun	nr	nr	nr
Gemini	61	1	flx	20 mg/kg	56 d	1	ki-67	nr	nr	nr
Gemini	61	2	flx	20 mg/kg	56 d	1	neun	nr	nr	nr
Gemini	61	3	flx	20 mg/kg	56 d	1	bdnf	nr	nr	nr
Manual	62	1	cit	15 mg/kg	35 d	1	ki-67	nr	nr	nr
GPT	62	1	nr	nr	nr	nr	ki-67	nr	nr	nr
GPT	62	2	cit	15 mg/kg	5 d	1	ki-67	nr	nr	nr
GPT	62	3	nr	nr	nr	nr	ki-67	nr	nr	nr
Gemini	62	1	cit	15 mg/kg	35 d	1	ki-67	nr	nr	nr

NR: não reportado; Ami: amitriptilina; Cit: citalopram; Des: desipramina; Esc: escitalopram; Flu: fluvoxamina; Flx: fluoxetina; Imi: imipramina; Prx: paroxetina. *Elaborado pela autora (2025).*

5.2.3 Extração de variáveis para o *desfecho*

A **Tabela 6** apresenta uma comparação entre as variáveis relacionadas ao modelo animal, i.e., **protocolos de estresse** e **testes comportamentais** extraídos pela forma tradicional (manual) vs. ChatGPT e Gemini.

Tabela 6. Variáveis para o desfecho (O) das extrações Manual, GPT e Gemini

Origem	ID	Estudo	Protocolo de Estresse	Teste Comportamental
Manual	1	1	cums	epm, fst, oft
GPT	1	1	cums	nr
GPT	1	2	cums	nr
GPT	1	3	cums	nr
Gemini	1	1	cums	epm, fst, oft
Manual	12	1	nr	mwm
GPT	12	1	nr	mwm
GPT	12	2	nr	mwm
Gemini	12	1	mwm	mwm
Gemini	12	2	mwm	mwm
Manual	15	1	nr	oft
Manual	15	2	nr	oft
Manual	15	3	nr	oft
Manual	15	4	nr	oft
Manual	15	5	nr	oft
Manual	15	6	nr	oft
Manual	15	7	nr	oft
Manual	15	8	nr	oft
GPT	15	1	nr	nr
GPT	15	2	nr	nr
GPT	15	3	nr	oft
GPT	15	4	nr	oft
GPT	15	5	nr	oft
GPT	15	6	nr	oft
GPT	15	7	nr	oft
GPT	15	8	nr	oft
GPT	15	9	nr	oft
GPT	15	10	nr	oft
Gemini	15	1	nr	oft
Gemini	15	2	nr	oft
Gemini	15	3	nr	oft

Gemini	15	4	nr	oft
Manual	16	1	cus	nsf, oft, tst
GPT	16	1	cus	nsf, oft, tst
Gemini	16	1	cus	nsf, oft, tst
Manual	18	1	cums	nsf, spl
Manual	18	2	cums	nsf, spl
GPT	18	1	cums	act, nsf, spl
Gemini	18	1	cums	nsf, oft, spl
Gemini	18	2	cums	nsf, oft, spl
Manual	19	1	srt	fst, mwm, oft
Manual	19	2	srt	fst, mwm, oft
GPT	19	1	nr	fst, oft
GPT	19	2	nr	mwm
GPT	19	3	nr	mwm
Gemini	19	1	nr	mwm
Gemini	19	2	nr	mwm
Gemini	19	3	nr	mwm
Gemini	19	4	nr	mwm
Gemini	19	5	nr	mwm
Gemini	19	6	nr	mwm
Gemini	19	7	nr	mwm
Gemini	19	8	nr	mwm
Gemini	19	9	nr	mwm
Gemini	19	10	nr	mwm
Gemini	19	11	nr	mwm
Gemini	19	12	nr	mwm
Manual	21	1	cums	epm, nest, oft, spl, tst
GPT	21	1	cums	epm, nest, oft, spl, tst
Gemini	21	1	cums	epm, nest, oft, spl, tst
Manual	25	1	cums	epm, spt, tst
GPT	25	1	cums	epm, fst, oft, spt, tst
GPT	25	2	cums	epm, fst, oft, spt, tst
Gemini	25	1	cums	epm, spt, tst
Gemini	25	2	cums	epm, spt, tst
Manual	28	1	cums	fst
GPT	28	1	cums	fst
Gemini	28	1	cums	fst
Manual	30	1	cums	oft, spt, tst
GPT	30	1	cums	oft, spt, tst
GPT	30	2	cums	oft, spt, tst
Gemini	30	1	cums	oft, spt, tst
Gemini	30	2	cums	oft, spt, tst
Gemini	30	3	cums	oft, spt, tst
Manual	31	1	cums	fst, nsf, oft, spt, tst
GPT	31	1	cums	fst, nsf, spt, tst
Gemini	31	1	cums	fst, nsf, spt, tst
Manual	32	1	nr	fst, tst
Manual	32	2	nr	fst, tst
GPT	32	1	nr	nr
Gemini	32	1	nr	fst, tst
Gemini	32	2	nr	fst, tst
Gemini	32	3	nr	fst, tst
Gemini	32	4	nr	fst, tst
Manual	39	1	wirs	fst, utt
Manual	39	2	wirs	fst, utt
GPT	39	1	wirs	fst, oft
GPT	39	2	wirs	fst, oft
Gemini	39	1	wirs	fcf, fst, oft
Gemini	39	2	wirs	fcf, fst, oft
Gemini	39	3	wirs	nr
Gemini	39	4	wirs	nr
Gemini	39	5	wirs	nr
Gemini	39	6	wirs	nr
Manual	40	1	nr	fst, oft, spt, tst
GPT	40	1	lps	nr
Gemini	40	1	lps	fst, oft, spt, tst

Manual	43	1	cums	oft
GPT	43	1	cus	oft
Gemini	43	1	cus	oft
Gemini	43	2	cus	oft
Manual	45	1	cums	oft, spt
GPT	45	1	cums, mcao, sir	bwt, cyl, gst, mwm, oft, spt
GPT	45	2	cums, mcao, sir	bwt, cyl, gst, mwm, oft, spt
Gemini	45	1	cums, mcao, sir	bwt, cyl, gst, mwm, oft, spt
Gemini	45	2	cums, mcao, sir	bwt, cyl, gst, mwm, oft, spt
Manual	48	1	cus	fst, sbx, tst
Manual	48	2	cus	fst, sbx, tst
GPT	48	1	cus	fst, sbx, tst
GPT	48	2	cus	fst, sbx, tst
GPT	48	3	cus	fst, sbx, tst
Gemini	48	1	cus	fst, sbx, tst
Gemini	48	2	cus	fst, sbx, tst
Gemini	48	3	cus	fst, sbx, tst
Gemini	48	4	cus	fst, sbx, tst
Manual	52	1	nr	fcf
Manual	52	2	nr	fcf
Manual	52	3	nr	fcf
Manual	52	4	nr	fcf
Manual	52	5	nr	fcf
Manual	52	6	nr	fcf
Manual	52	7	nr	fcf
Manual	52	8	nr	fcf
Manual	52	9	nr	fcf
Manual	52	10	nr	fcf
Manual	52	11	nr	fcf
Manual	52	12	nr	fcf
Manual	52	13	nr	fcf
Manual	52	14	nr	fcf
Manual	52	15	nr	fcf
Manual	52	16	nr	fcf
Manual	52	17	nr	fcf
Manual	52	18	nr	fcf
Manual	52	19	nr	fcf
Manual	52	20	nr	fcf
Manual	52	21	nr	fcf
Manual	52	22	nr	fcf
Manual	52	23	nr	fcf
Manual	52	24	nr	fcf
Manual	52	25	nr	fcf
Manual	52	26	nr	fcf
Manual	52	27	nr	fcf
Manual	52	28	nr	fcf
Manual	52	29	nr	fcf
Manual	52	30	nr	fcf
Manual	52	31	nr	fcf
Manual	52	32	nr	fcf
Manual	52	33	nr	fcf
Manual	52	34	nr	fcf
Manual	52	35	nr	fcf
Manual	52	36	nr	fcf
Manual	52	37	nr	fcf
Manual	52	38	nr	fcf
Manual	52	39	nr	fcf
Manual	52	40	nr	fcf
Manual	52	41	nr	fcf
Manual	52	42	nr	fcf
Manual	52	43	nr	fcf
Manual	52	44	nr	fcf
GPT	52	1	nr	fcf
GPT	52	2	nr	fcf
Gemini	52	1	nr	nr
Gemini	52	2	nr	nr
Gemini	52	3	nr	nr

Gemini	52	4	nr	nr
Gemini	52	5	nr	fcp
Gemini	52	6	nr	fcp
Gemini	52	7	nr	fcp
Gemini	52	8	nr	fcp
Gemini	52	9	nr	fcp
Gemini	52	10	nr	fcp
Gemini	52	11	nr	fcp
Gemini	52	12	nr	fcp
Manual	53	1	cums	fst, oft, spt, tst
GPT	53	1	cums	fst, oft, spt, tst
Gemini	53	1	cums	fst, spt, tst
Manual	54	1	nr	tst
Manual	54	2	nr	tst
Manual	54	3	nr	tst
Manual	54	4	nr	tst
Manual	54	5	nr	tst
Manual	54	6	nr	tst
Manual	54	7	nr	tst
Manual	54	8	nr	tst
Manual	54	9	nr	tst
Manual	54	10	nr	tst
Manual	54	11	nr	tst
Manual	54	12	nr	tst
Manual	54	13	nr	tst
Manual	54	14	nr	tst
Manual	54	15	nr	tst
GPT	54	1	nr	tst
GPT	54	2	nr	tst
GPT	54	3	nr	tst
GPT	54	4	nr	tst
GPT	54	5	nr	tst
GPT	54	6	nr	tst
Gemini	54	1	nr	tst
Gemini	54	2	nr	tst
Gemini	54	3	nr	tst
Gemini	54	4	nr	tst
Gemini	54	5	nr	nr
Gemini	54	6	nr	nr
Manual	55	1	nr	fst, tst
Manual	55	2	cums	oft, spt
GPT	55	1	nr	nr
Gemini	55	1	cums	oft, spt
Manual	58	1	nr	tst
GPT	58	1	nr	fst
Gemini	58	1	nr	fst
Manual	59	1	ms	fst
GPT	59	1	els	fst
Gemini	59	1	ms	fst
Gemini	59	2	ms	fst
Manual	60	1	nr	rot, spt, tst
GPT	60	1	rotenone	rot, spt, tst
Gemini	60	1	rotenone	rot, spt, tst
Gemini	60	2	rotenone	rot, spt, tst
Manual	61	1	cums	epm, fst, nsf, sit
GPT	61	1	cums	epm, nsf, other
GPT	61	2	cums	epm, nsf, other
Gemini	61	1	cums	epm, fst, nsf, sit
Gemini	61	2	cums	epm, fst, nsf, sit
Gemini	61	3	cums	epm, fst, nsf, sit
Manual	62	1	cms	fcp, fst, spt
GPT	62	1	cms	fcp, fst, spt
GPT	62	2	cms	fcp, fst, spt
GPT	62	3	cms	fcp, fst, spt
Gemini	62	1	cms	fcp, fst, spt

NR: não reportado; ACT: Teste de Actímetro; BWT: Teste de Travessia na Barra; CMS: Estresse Crônico

Moderado; CUMS: Estresse Crônico Moderado Imprevisível; CUS: Estresse Crônico Imprevisível; CYL: Teste do Cilindro; ELS: Estresse Precoces; EPM: Teste de Labirinto em Cruz Elevado; FCP: Teste de Paradigma de Condicionamento do Medo; FST: Teste de Nado Forçado; GST: Teste de Força de Preensão; LPS: Modelo de Depressão Induzida por LPS; MWM: Teste do Labirinto Aquático de Morris; MS: Estresse de Separação Maternal; NSF: Teste de Alimentação Suprimida pela Novidade; OFT: Teste de Campo Aberto; ROT Teste do Rotarod; SBX: Teste da Caixa de Transição; SIR: Estresse por Isolamento Social; SIT: Teste de Interação Social; SPL: Teste do Borriço; SPT: Teste de Preferência por Sacarose; TST: Teste de Suspensão pela Cauda; UTT: Teste de Resposta Inata ao Som; WIRS: Estresse por Contenção e Imersão em Água. *Elaborado pela autora (2025).*

5.2.4 Comparação estrutural de estudos extraídos por artigo pré e pós-alinhamento

A simples comparação linha-a-linha entre as planilhas se mostrou insuficiente para a validação estatística, pois essa limitação decorre da não-linearidade intrínseca à saída (*output*) dos LLMs. Foi observado que os modelos de IA, mesmo quando capazes de extrair corretamente o conteúdo de múltiplos experimentos, frequentemente os listava em uma ordem distinta daquela adotada pela avaliadora na extração manual. Isso pode ser identificado, por exemplo, no artigo de ID 15 (**Tabela 4, 5 e 6**), no qual o modelo inverteu a ordem de apresentação dos grupos experimentais, fazendo com que o “Estudo 1” da extração manual correspondesse, na verdade, à segunda linha da extração automatizada (i.e., “Estudo 2”). Em uma verificação linear tradicional, a inversão de estudos resultaria em falsos erros de concordância. Dessa forma, como última etapa do processamento dos dados antes das análises de desempenho, foi realizado um alinhamento estrutural dos *dataframes*. A abordagem de alinhamento foi estatisticamente confirmada pela taxa de reordenamento, pois o algoritmo *best-fit* detectou que o par correto não se encontrava na posição linear correspondente em 18 casos para a planilha ChatGPT e 31 casos para a planilha Gemini (**Tabela 7**), executando-se assim o pareamento. Para além dos pares correspondentes, a análise estrutural pós-alinhamento permitiu contabilizar 54 e 46 estudos para o ChatGPT e para o Gemini, respectivamente, que estavam presentes no Manual e que os modelos falharam em capturar, conceituando os falsos negativos estruturais (**Tabela 7**). Adicionalmente, o ChatGPT e o Gemini geraram 13 e 28 linhas a mais, respectivamente, que não possuíam correspondência com nenhum estudo do Manual, conceituando os falsos positivos estruturais (**Tabela 7**).

Tabela 7. Pareamento dos estudos e identificação de omissões e alucinações

	Pares alinhados (<i>best-fit</i>)	Falsos negativos estruturais	Falsos positivos estruturais
ChatGPT (gratuito)	18	54	13
Gemini Pro 2.5	31	46	28

Elaborado pelo autora (2025).

A comparação estrutural a nível pré-alinhamento (**Figura 8a**) entre as extrações realizadas usando o ChatGPT, demonstrou que apenas 42,30% dos artigos incluídos apresentaram o mesmo número de estudos quando comparados à extração manual. Ou seja, o ChatGPT só foi capaz de recuperar o mesmo número de estudos para 11 artigos, sendo estes os IDs 16, 21, 28, 31, 39, 40, 43, 53, 58, 59 e 60. Já em nível pós-alinhamento (**Figura 8b**), a composição final dos pares estabelecidos pelo algoritmo *best-fit* gráfico confirma a hipótese de que a divergência volumétrica não seria apenas numérica, mas de conteúdo. A figura mostra os estudos categorizados em pares correspondentes, dos quais os falsos negativos representam os dados extraídos manualmente, mas não identificados pelo ChatGPT, enquanto os falsos positivos representam os dados extraídos pelo ChatGPT, mas não manualmente. Em ambos os momentos (pré e pós-alinhamento), o artigo identificado como 52 apresentou grande divergência entre as abordagens usadas para a extração de dados (**Figura 8a**), pois a extração manual identificou 44 estudos, porém o ChatGPT recuperou um número ínfimo de apenas 2 estudos. O inverso foi observado para o artigo 15, em que o ChatGPT identificou 2 estudos a mais que a extração manual. Conforme observado na **Figura 8b**, os falsos negativos foram mais frequentes nos artigos 18, 32, 52, 54 e 55, sendo particularmente predominante nos artigos 52 e 54. Já os falsos positivos estiveram presentes nos artigos 1, 12, 15, 19, 25, 30, 45, 48, 61 e 62.

A comparação estrutural a nível pré-alinhamento (**Figura 9a**) entre as extrações realizadas usando o Gemini, demonstrou que 38,46% dos artigos incluídos apresentaram o mesmo número de estudos quando comparados à extração manual. Os 10 artigos com números iguais de estudos foram os IDs 1, 16, 18, 21, 28, 31, 40, 53, 58 e 62. Semelhante ao ChatGPT, o Gemini também não foi capaz de recuperar o número correto de estudos na maioria dos casos. Igualmente com o artigo 52, o Gemini falhou em encontrar o número de estudos encontrados na extração manual, retornando apenas 12 estudos dos 44 extraídos pela avaliadora. O Gemini também apresentou tendência a “inventar” estudos como o ChatGPT, evento observado no artigo 19, onde o Gemini registrou 10 estudos a mais do que a extração manual. Já em nível pós-alinhamento (**Figura 9b**), foi possível observar que os falsos negativos foram mais frequentes nos artigos identificados pelos IDs 15, 52, 54 e 55, enquanto os falsos positivos estiveram presentes nos artigos 12, 19, 25, 30, 32, 39, 43, 45, 48, 59, 60 e 61. Com base nessa comparação, é possível observar que o Gemini, assim como o ChatGPT, apresenta falhas na identificação e classificação correta das informações.

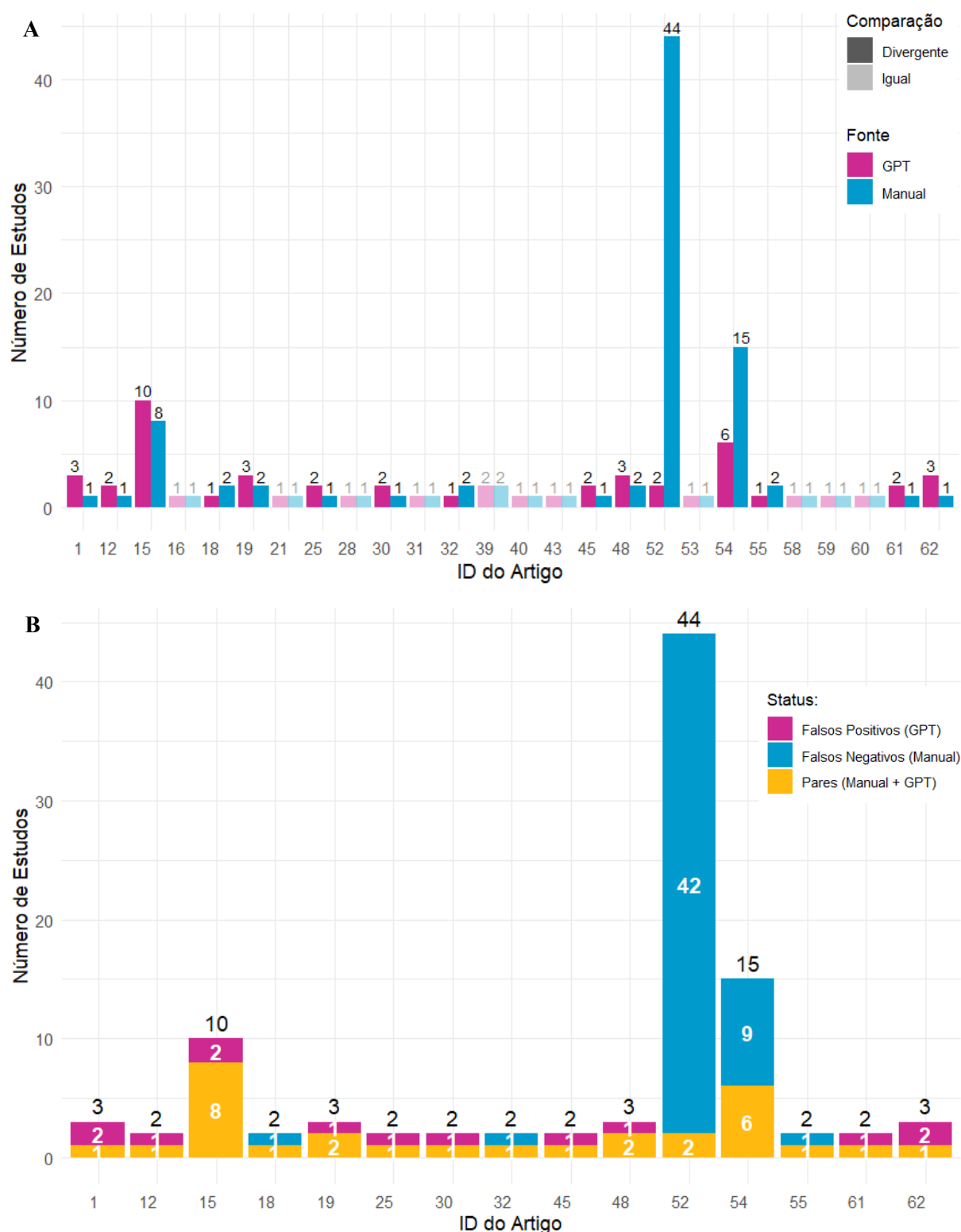


Figura 8. Comparação estrutural de estudos extraídos por artigo antes (a) e após o alinhamento (b) com o ChatGPT. A) Contagem de estudos da extração manual (azul) e pelo modelo GPT (rosa) para cada artigo (ID). Barras sólidas indicam contagens diferentes e barras translúcidas indicam concordância no volume de estudos extraídos. B) Resultados do algoritmo de alinhamento com composição dos artigos divergentes, indicando os pares (verde), falsos negativos (azul) e falsos positivos (rosa). *Elaborado pela autora (2025).*

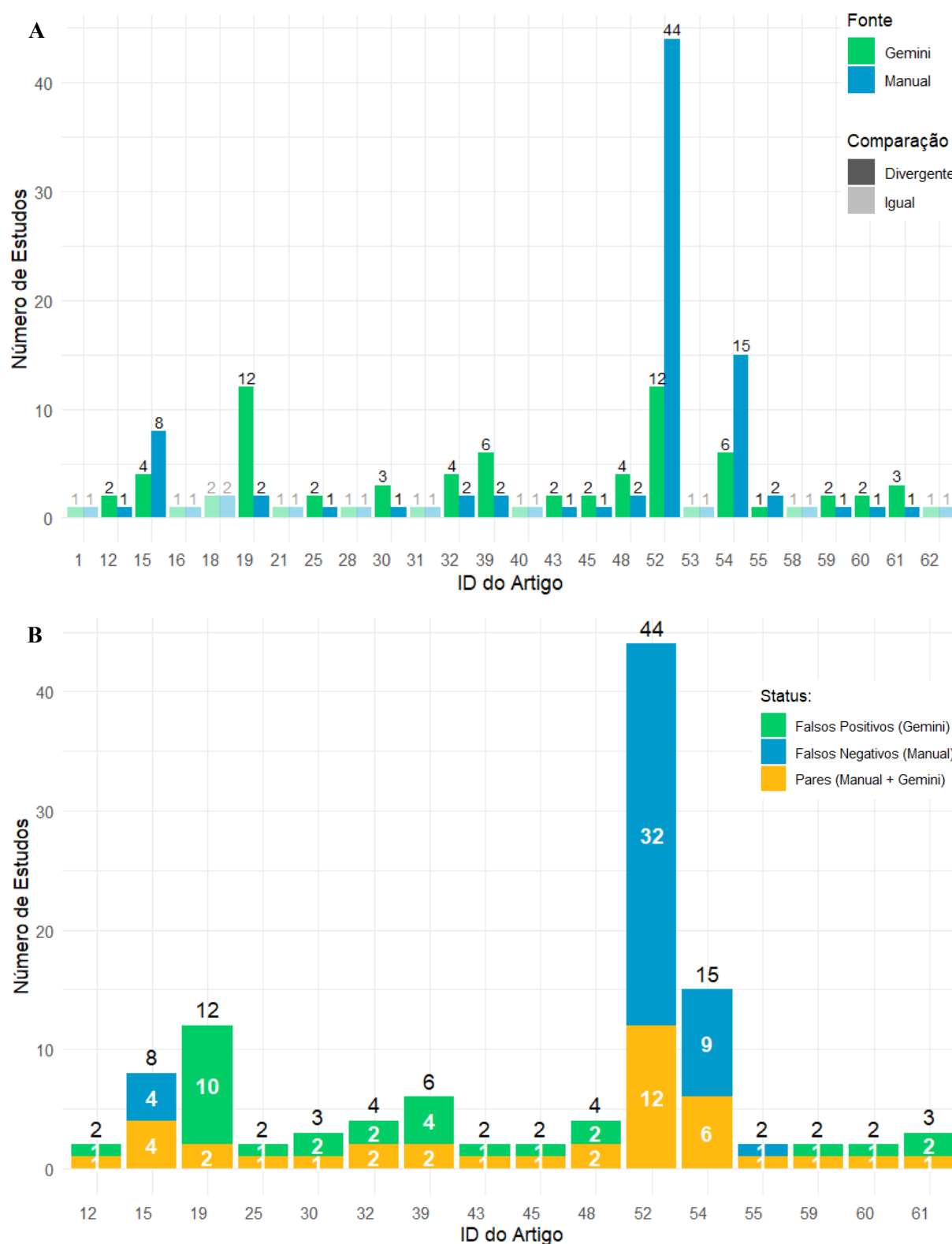


Figura 9. Comparação estrutural de estudos extraídos por artigo antes (a) e após o alinhamento (b) com o Gemini. A) Contagem de estudos da extração manual (azul) e pelo modelo Gemini (verde) para cada artigo (ID). Barras sólidas indicam contagens diferentes e barras translúcidas indicam concordância no volume de estudos extraídos. B) Resultados do algoritmo de alinhamento com composição dos artigos divergentes, indicando os pares (laranja), falsos negativos (azul) e falsos positivos (verde). *Elaborado pela autora (2025).*

5.3 Análises estatísticas e métricas de desempenho das ferramentas automatizadas

Com os dados tratados e alinhados, tornou-se possível empregar adequadamente as análises estatísticas para estimar a concordância entre a extração automatizada *versus* extração manual. O tempo decorrido para o desenvolvimento dos códigos para o processamento dos dados e para as análises exigiu um trabalho de aproximadamente dez meses. Todos os códigos estão disponíveis no [OSF](#) (em “Files/Extração/Dissertação_Hamze_2026/Scripts”).

Visto que utilizar dados não pareados consequentemente violaria os pressupostos estatísticos dos testes, como o Kappa e as matrizes de confusão, para análise de *tokens* se utilizou o *dataframe* de dados tratados (seção 5.3.1) e para as análises de concordância se utilizou o *dataframe* de dados alinhados (seção 5.3.2).

5.3.1 Métricas de recuperação para os dados extraídos pelos modelos de IA

Métricas de precisão, sensibilidade (ou revocação) e F_1 -score foram calculadas para todas as variáveis presentes nos artigos incluídos (**Figura 10**). As métricas de desempenho para a extração com o Gemini foram maiores do que para a extração com o ChatGPT, exceto para a variável espécie.

Para o ChatGPT, o gráfico de desempenho global revela que as variáveis espécie, marcador de neurogênese e injeções de BrdU alcançaram F_1 -score $>80\%$. As variáveis tempo de sobrevivência, dose de BrdU, protocolo de estresse, teste comportamental, linhagem, antidepressivo e sexo alcançaram um desempenho de 60 a 80%. As variáveis com o pior desempenho ($<40\%$) foram dose de antidepressivo, duração do tratamento e número de injeções de antidepressivo. Para a variável espécie, o F_1 -score foi de 98%, sendo considerada como “quase perfeita”. A variável número de injeções de antidepressivo foi classificada como pior desempenho, devido ao F_1 -score de 36%.

Os dados extraídos pelo Gemini para espécie, antidepressivo, sexo, linhagem, teste comportamental, injeções de antidepressivo, dose de antidepressivo e marcador de neurogênese alcançaram F_1 -score superiores a 80%. As variáveis com desempenho intermediário (60 a 80%) foram tempo de sobrevivência, dose de BrdU, número de injeções de BrdU, protocolos de estresse e tempo de tratamento. Nenhuma variável apresentou desempenho insatisfatório ($<40\%$). Para a variável espécie, o F_1 -score foi de 98%, sendo considerada como “quase perfeita”. A variável tempo de tratamento do antidepressivo foi classificada como pior desempenho, devido ao F_1 -score de 64%.

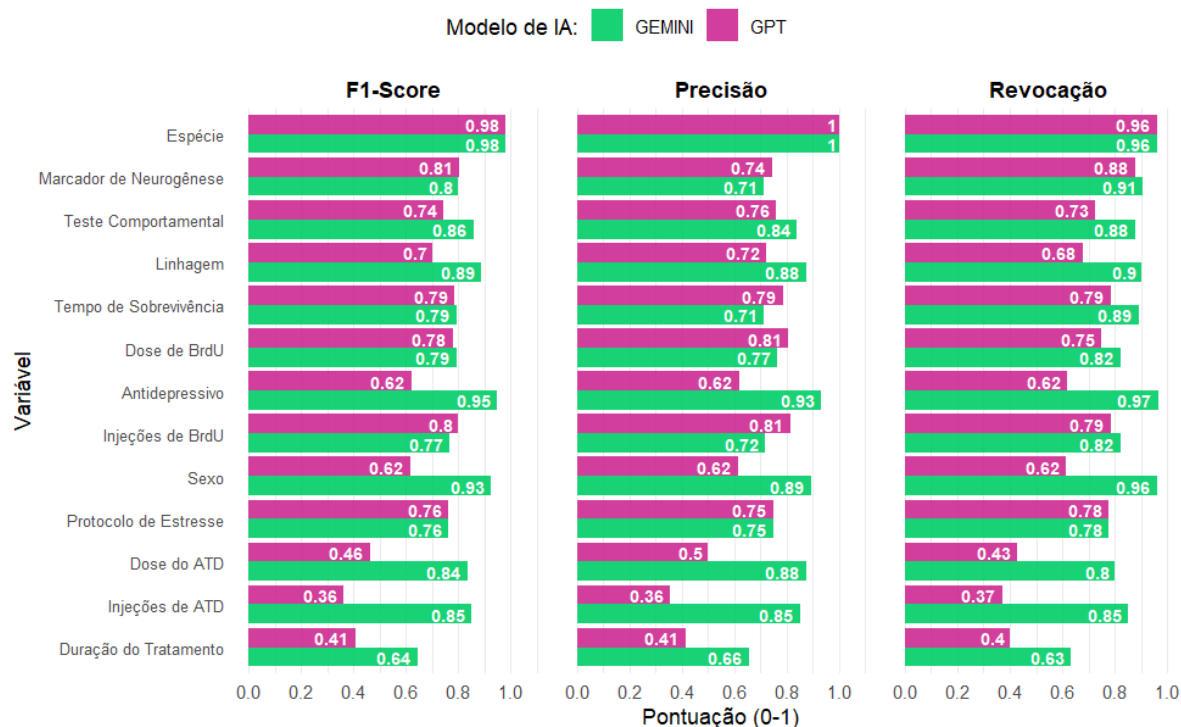


Figura 10. Análise de desempenho global por variável da extração automatizada do Gemini vs. ChatGPT. O gráfico de barras apresenta as métricas de precisão, sensibilidade e F_1 -score calculadas a nível de artigos. A análise mostra que o Gemini (verde) apresenta um melhor desempenho do que o GPT (rosa) para extrair os dados de interesse. *Elaborado pela autora (2025).*

5.3.2 Métricas de concordância para os dados extraídos pelos modelos de IA

As análises de concordâncias pelo Kappa de Cohen foram calculadas para quantificar o nível de concordância entre o humano e o ChatGPT ou Gemini para variáveis categóricas de rótulo único (*single-label*).

Para o ChatGPT (**Figura 11a**), as concordâncias variaram de concordância perfeita para espécie (100%) ou quase perfeita para marcador de neurogênese (82%), substancial para linhagem (76%), dose de BrdU (73%), injeções de BrdU (70%), tipo de antidepressivo (69%) e tempo de sobrevivência (68%), moderada para duração do tratamento (48%), e discreta para sexo (11%) e injeções de antidepressivo (6%). As análises de concordância para o Gemini (**Figura 11b**) demonstraram concordância perfeita (100%) ou quase perfeita (>80%) para as variáveis espécies (100%), linhagens (98%), tipo de antidepressivo (97%), marcador de neurogênese (90%) e dose do antidepressivo (88%) (Figura 24). Também obtiveram concordância substancial (>60%) os dados extraídos para os dados extraídos para o protocolo de BrdU, incluindo tempo de sobrevivência (67%) e número de injeções (64%). Concordâncias moderadas (>40%) foram obtidas nas variáveis dose de BrdU (59%) e duração do tratamento (50%). Concordâncias mais baixas (<40%) foram registradas para a variável sexo (28%) e

número de injeções de antidepressivo (12%).

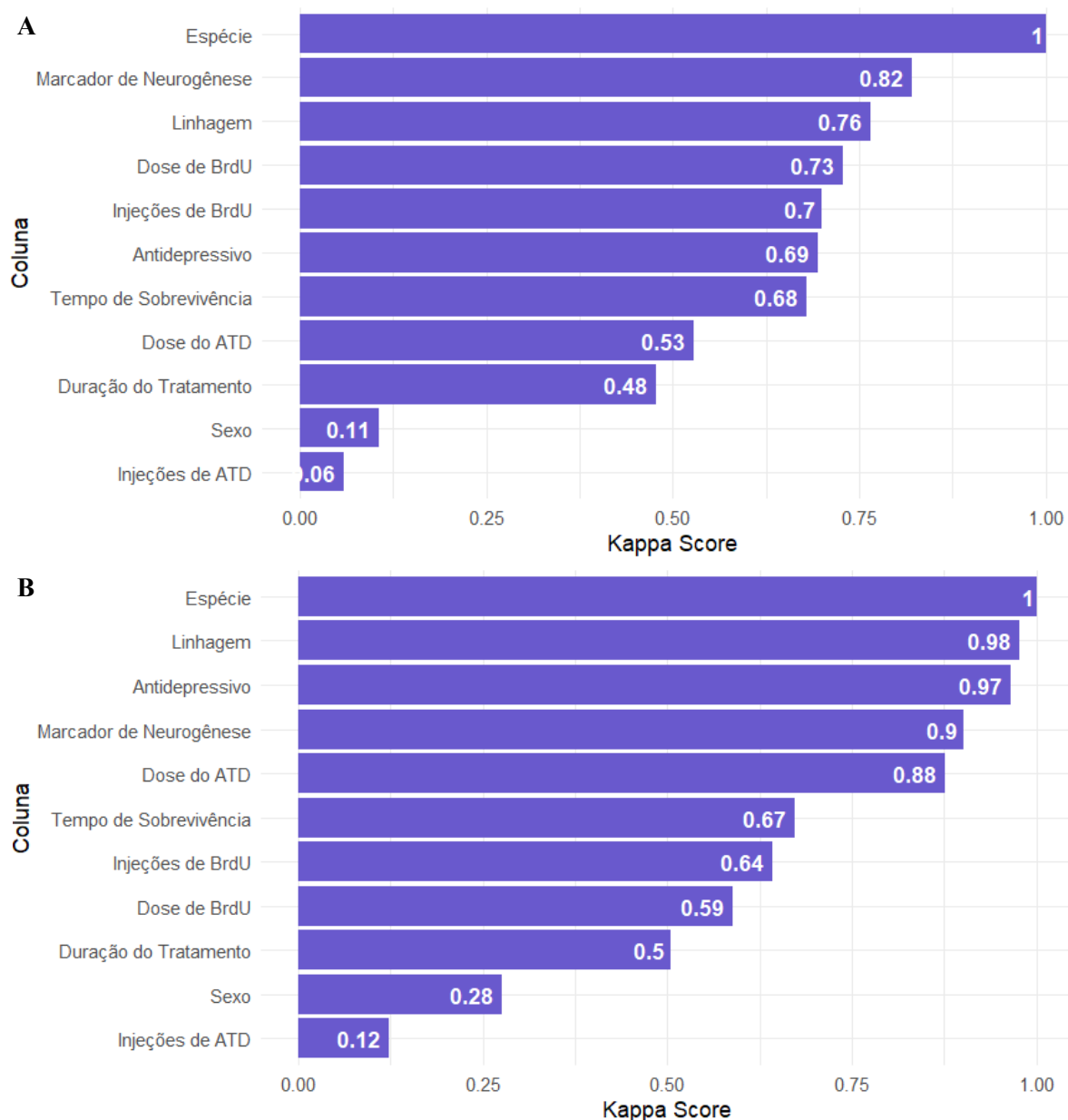


Figura 11. Análises de concordâncias para variáveis *single-label* da extração automatizada com o ChatGPT (a) e Gemini (b). Para os estudos alinhados, os valores de Kappa foram calculados entre a extração manual e a extração pelos modelos ChatGPT ou Gemini. A métrica isola a concordância real ajustando os acertos à concordância ao acaso. *Elaborado pela autora (2025).*

As matrizes de confusão para as variáveis *single-label* foram construídas para permitir uma melhor visualização dos dados extraídos (real vs. predito) e são apresentadas de acordo com os elementos do acrônimo PICO (população, intervenção e desfecho). Para as extrações pelo ChatGPT, as matrizes construídas para as variáveis espécie, sexo e linhagem são apresentadas na **Figura 12**. Para a espécie (**Figura 12a**), o ChatGPT obteve apenas 34 acertos para camundongos (“mice”) e 54 vezes o classificou como “nr”. Para o sexo (**Figura 12b**), o

ChatGPT obteve 30 acertos para macho, mas o classificou como “nr” 22 vezes. As linhagens (Figura 12c) também sofreram maior classificação incorreta, onde o maior acerto foi para a linhagem “swiss” com 9 acertos absolutos (i.e., sem perder nenhum *token* “swiss” para “nr”), mas outras linhagens como “euploid” e “ts65dn” representaram os maiores erros e obtiveram 22 e 20 perdas para “nr”, respectivamente.

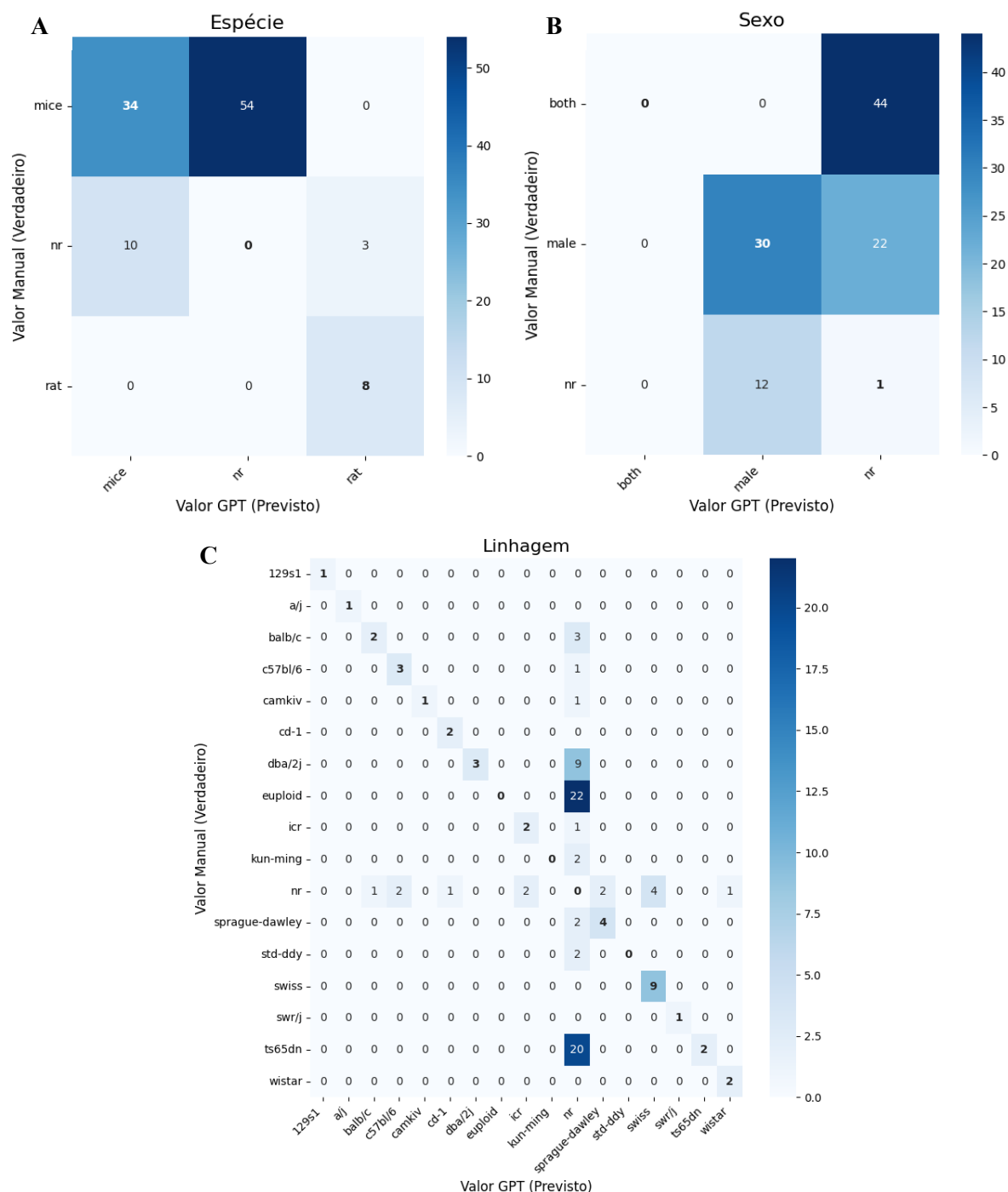


Figura 12. Matriz de confusão para a espécie (a), sexo (b) e linhagem (c) para os dados extraídos no GPT. O eixo y (vertical) representa a classificação do padrão-ouro ou referência de valor verdadeiro (Manual) e o eixo x (horizontal) a predição pelo modelo (GPT). A diagonal principal (negrito) indica acertos, e a coloração mais intensa das células indica a quantidade de *tokens* contados. NR: não reportado. Elaborado pela autora (2025).

As matrizes para as extrações pelo Gemini são apresentadas na **Figura 13**. Para a espécie (**Figura 13a**), o Gemini identificou corretamente 42 estudos que utilizaram camundongos e 8 estudos que utilizaram ratos. O erro mais frequente foi “nr” em 23 estudos. O Gemini também identificou corretamente o sexo em 38 estudos que utilizaram machos, mas falhou (n=0) quando o estudo menciona o uso de machos e fêmeas concomitantemente, atribuindo “nr” (**Figura 13b**). A linhagem “ts65dn” foi a mais bem classificada pelo modelo com 12 acertos, mas errou a linhagem “euploid” trocando por “nr” 22 vezes (**Figura 13c**).

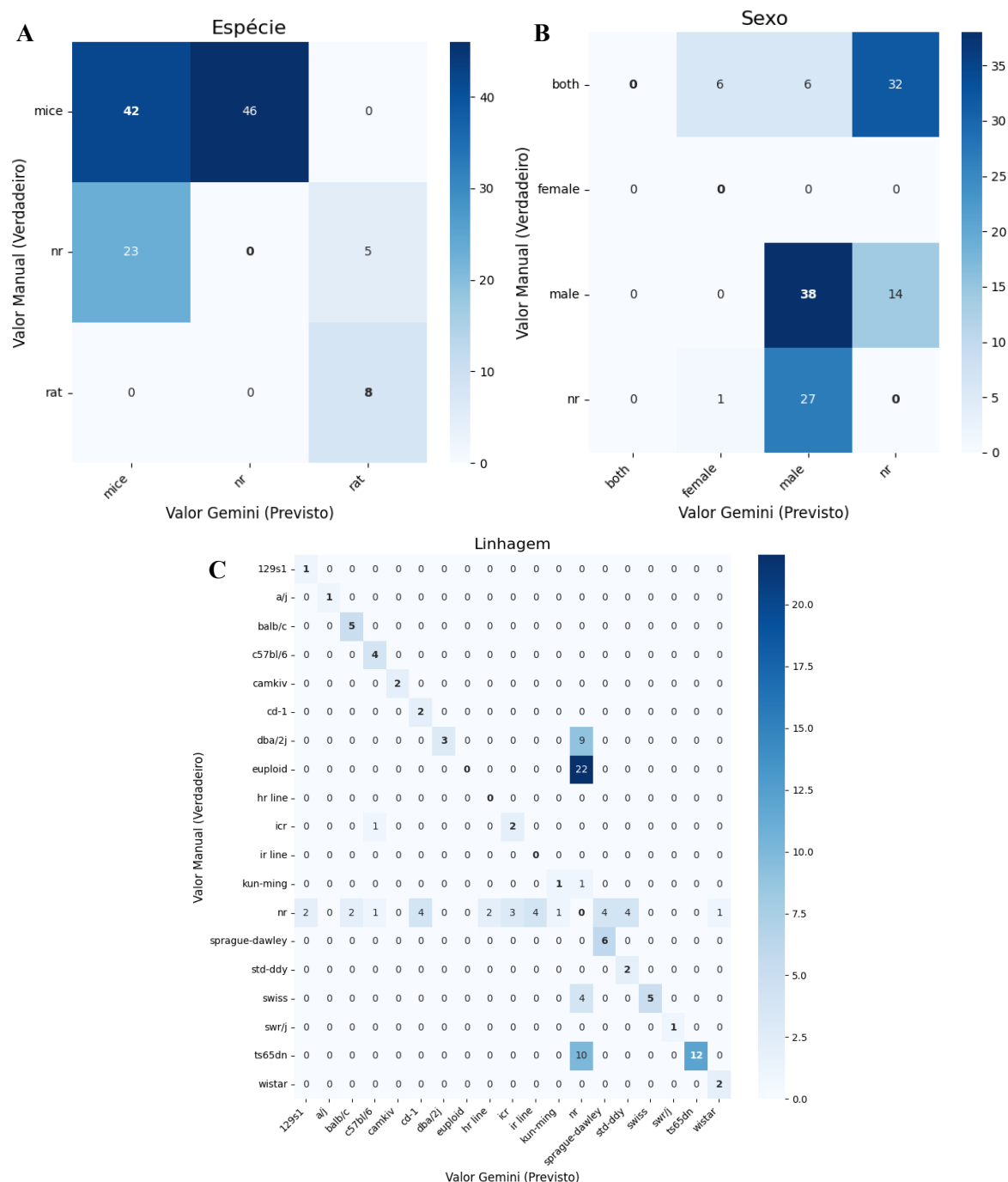


Figura 13. Matriz de confusão para a espécie (a), sexo (b) e linhagem (c) para os dados extraídos no Gemini. O eixo y (vertical) representa a classificação do padrão-ouro ou referência de valor verdadeiro (Manual) e o eixo

x (horizontal) a predição pelo modelo (Gemini). A diagonal principal (negrito) indica acertos, e a coloração mais intensa das células indica a quantidade de *tokens* contados. NR: não reportado. *Elaborado pela autora (2025).*

Na **Figura 14** são apresentadas as matrizes construídas para a extração pelo ChatGPT para as variáveis tipo de antidepressivo, número de injeções, dose e duração do tratamento. Para a matriz construída para o tipo de antidepressivo (**Figura 14a**), é possível observar que o maior acerto do ChatGPT foi para “flx” com 19 acertos, contudo, o mesmo token também representa o maior erro com 53 perdas para “nr”. Para a matriz construída para a dose de antidepressivos (**Figura 14b**), o ChatGPT acertou pouco e errou muito, com 7 acertos para “nr” e 5 acertos para as doses de 10 e 18 mg/kg, mas, somando todos os valores, obteve um total de 69 erros de “nr”. Para a matriz de injeções de antidepressivo (**Figura 14c**), é possível observar que ChatGPT obteve mais acertos para “1” com 18 correspondências, ao mesmo tempo que perdeu esse exato valor para “nr” 70 vezes. Para a matriz da duração do tratamento (**Figura 14d**), é possível observar que ChatGPT também teve poucos acertos, com apenas 8 acertos de “21 day” e 7 acertos de “nr”, mas 67 erros ao todo de “nr”.

Na **Figura 15** são apresentadas as matrizes construídas para a extração pelo Gemini para as variáveis tipo de antidepressivo, número de injeções, dose e duração do tratamento. Para a matriz construída para o tipo de antidepressivo (**Figura 15a**), é possível observar que a fluoxetina, antidepressivo mais frequente, foi identificado e classificado corretamente em 31 estudos. Entretanto, 41 vezes o Gemini não foi capaz de identificar o fármaco. O inverso também foi observado, por exemplo, nos casos em que os dados extraídos manualmente relatam ausência de informação, mas o Gemini relata o uso de fluoxetina 17 vezes. Para a matriz da dose de antidepressivo (**Figura 15b**), o maior número de erros também foi em relação à incapacidade de o Gemini identificar a informação presente no texto. A presença de “nr” foi frequente com as doses de 5 e 10 mg/kg, com 18 erros para ambas. Adicionalmente, houve casos em que o modelo extraiu, por exemplo, o valor de 10 mg/kg, indicando 13 erros de um valor ausente na extração manual. Para a matriz construída para o número de injeções de antidepressivos (**Figura 15c**), a comparação entre os dados extraídos manualmente vs. Gemini indicou que mais da metade dos casos em que o artigo relatou uma única injeção, o modelo interpretou como informação não relatada “nr”. Finalmente, para a duração do tratamento (**Figura 15d**) ocorreu erro de “nr” em 32 estudos com protocolos que utilizaram 12 dias de tratamento. Os acertos foram limitados, com os maiores acertos sendo em “21 day” e “14 day” com apenas 10 e 6 acertos, respectivamente.

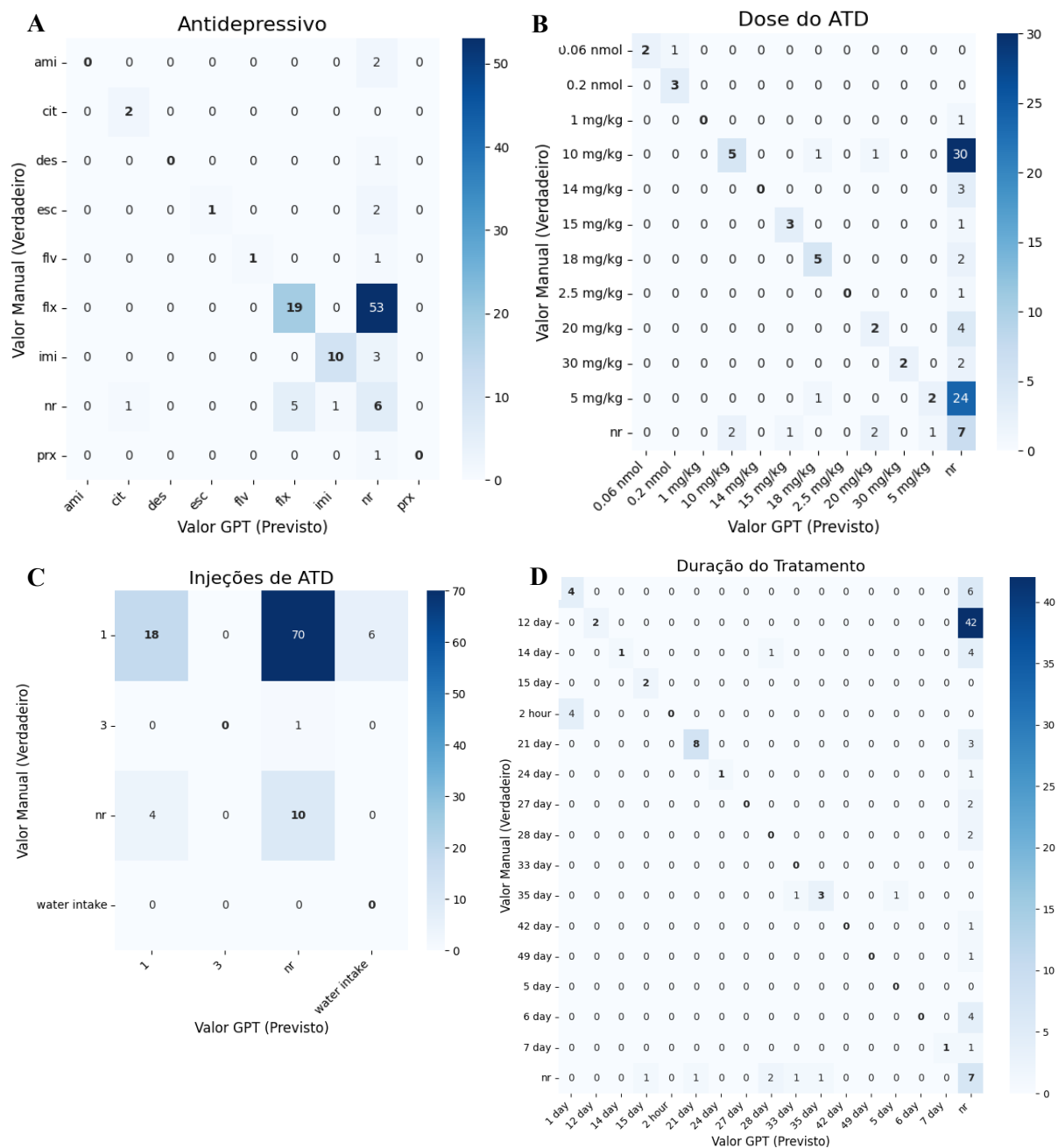


Figura 14. Matriz de confusão para antidepressivos (a), dose de antidepressivo (b), número de injeções de antidepressivo (c) e duração do tratamento (d) para os dados extraídos no ChatGPT. O eixo y (vertical) representa a classificação do padrão-ouro ou referência de valor verdadeiro (Manual) e o eixo x (horizontal) a predição pelo modelo (GPT). A diagonal principal (negrito) indica acertos, e a coloração mais intensa das células indica a quantidade de *tokens* contados. NR: não reportado; Water Intake: via oral; Ami: amitriptilina; Cit: citalopram; Des: desipramina; Esc: escitalopram; Flu: fluvoxamina; Flx: fluoxetina; Imi: imipramina; Prx: paroxetina. *Elaborado pela autora (2025).*

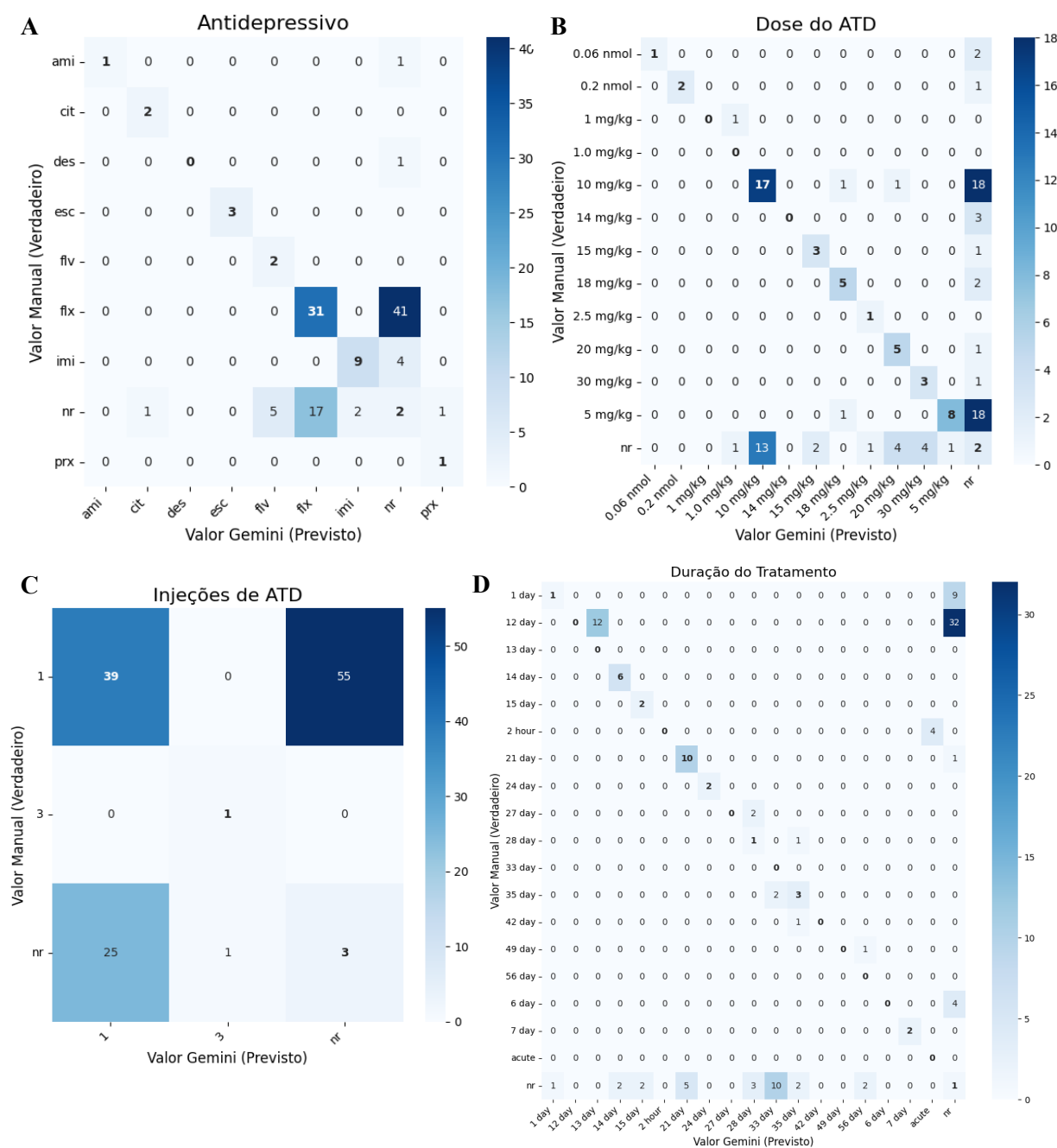


Figura 15. Matriz de confusão para antidepressivos (a), dose de antidepressivo (b), número de injeções de antidepressivo (c) e duração do tratamento (d) para os dados extraídos no Gemini. O eixo y (vertical) representa a classificação do padrão-ouro ou referência de valor verdadeiro (Manual) e o eixo x (horizontal) a predição pelo modelo (GPT). A diagonal principal (negrito) indica acertos, e a coloração mais intensa das células indica a quantidade de *tokens* contados. NR: não reportado; Ami: amitriptilina; Cit: citalopram; Des: desipramina; Esc: escitalopram; Flu: fluvoxamina; Flx: fluoxetina; Imi: imipramina; Prx: paroxetina. *Elaborado pela autora (2025).*

Para a matriz das variáveis de desfecho na extração pelo ChatGPT (**Figura 16**), o modelo classificou corretamente a maior parte dos marcadores BrdU, Ki-67 e DCX, com 18, 10 e 9 acertos, respectivamente (**Figura 16a**). Entretanto, o ChatGPT seguiu classificando como “nr” dados existentes no artigo. A matriz de confusão construída para as doses de BrdU

(Figura 16b) teve zero acertos para dose de 150 ug/g, mas 44 erros de “nr”. Além disso, teve 36 acertos de “nr”. A matriz de confusão construída para o número de injeções de BrdU (Figura 16c) mostra um erro sistemático para os estudos que relataram o uso de uma única injeção. A linha referente a uma injeção (“1”) mostra 51 perdas para “nr”, com apenas 6 acertos. Finalmente, para a matriz de confusão construída para o tempo de sobrevivência (Figura 16d), o ChatGPT falhou ao identificar a informação 36 vezes em protocolos de 30 dias (“30 day”), classificando-o como “nr”. Ao mesmo tempo, acertou 36 vezes ao classificar “nr” quando este também era o *token* identificado pela extração manual.

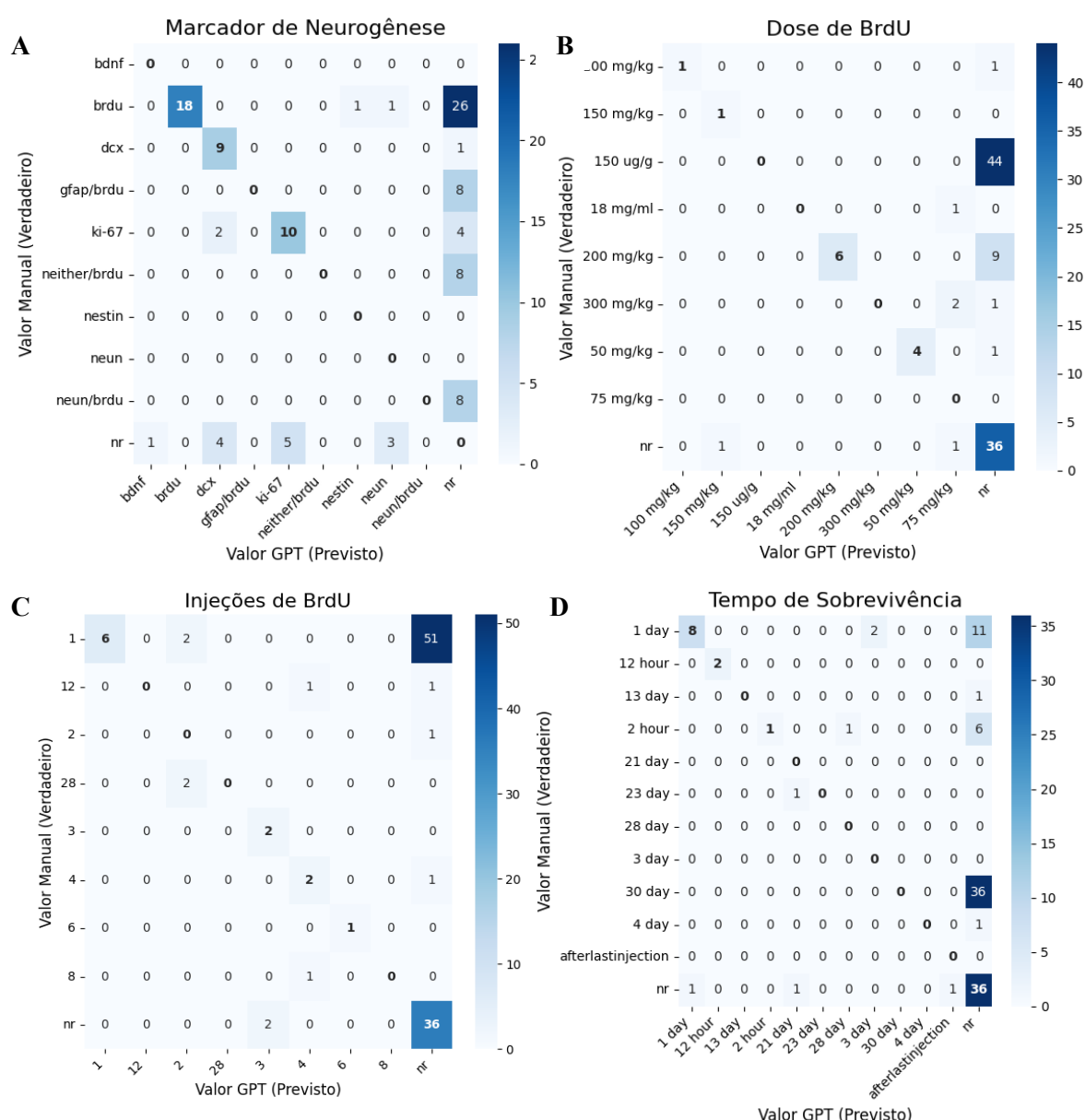


Figura 16. Matriz de confusão para marcador de neurogênese (a), dose de BrdU (b), número de injeções de BrdU (c) e tempo de sobrevivência (d) para os dados extraídos no GPT. O eixo y (vertical) representa a classificação do padrão-ouro ou referência de valor verdadeiro (Manual) e o eixo x (horizontal) a predição pelo modelo (GPT). A diagonal principal (negrito) indica acertos, e a coloração mais intensa das células indica a quantidade de *tokens* contados. NR: não reportado. *Elaborado pela autora (2025).*

Para a matriz das variáveis de desfecho na extração pelo Gemini (**Figura 17**), a matriz de confusão construída para os marcadores de neurogênese (**Figura 17a**) demonstrou que o erro mais frequente é “nr”. A matriz de confusão construída para as doses de BrdU (**Figura 17b**) teve 8 acertos para a dose de 150 ug/g, mas 36 perdas para “nr”. No entanto, houve 45 acertos quando a informação estava, de fato, ausente nos estudos. A matriz de confusão construída para o número de injeções de BrdU (**Figura 17c**) mostra um erro sistemático para os estudos que relataram o uso de uma única injeção, onde dados referentes a uma injeção mostram 45 perdas para “nr” e apenas 12 acertos. A matriz de confusão construída para o tempo de sobrevivência (**Figura 17d**), demonstrou que o modelo falhou 32 vezes ao identificar protocolos de 30 dias (“30 day”), classificando-os como “nr”.

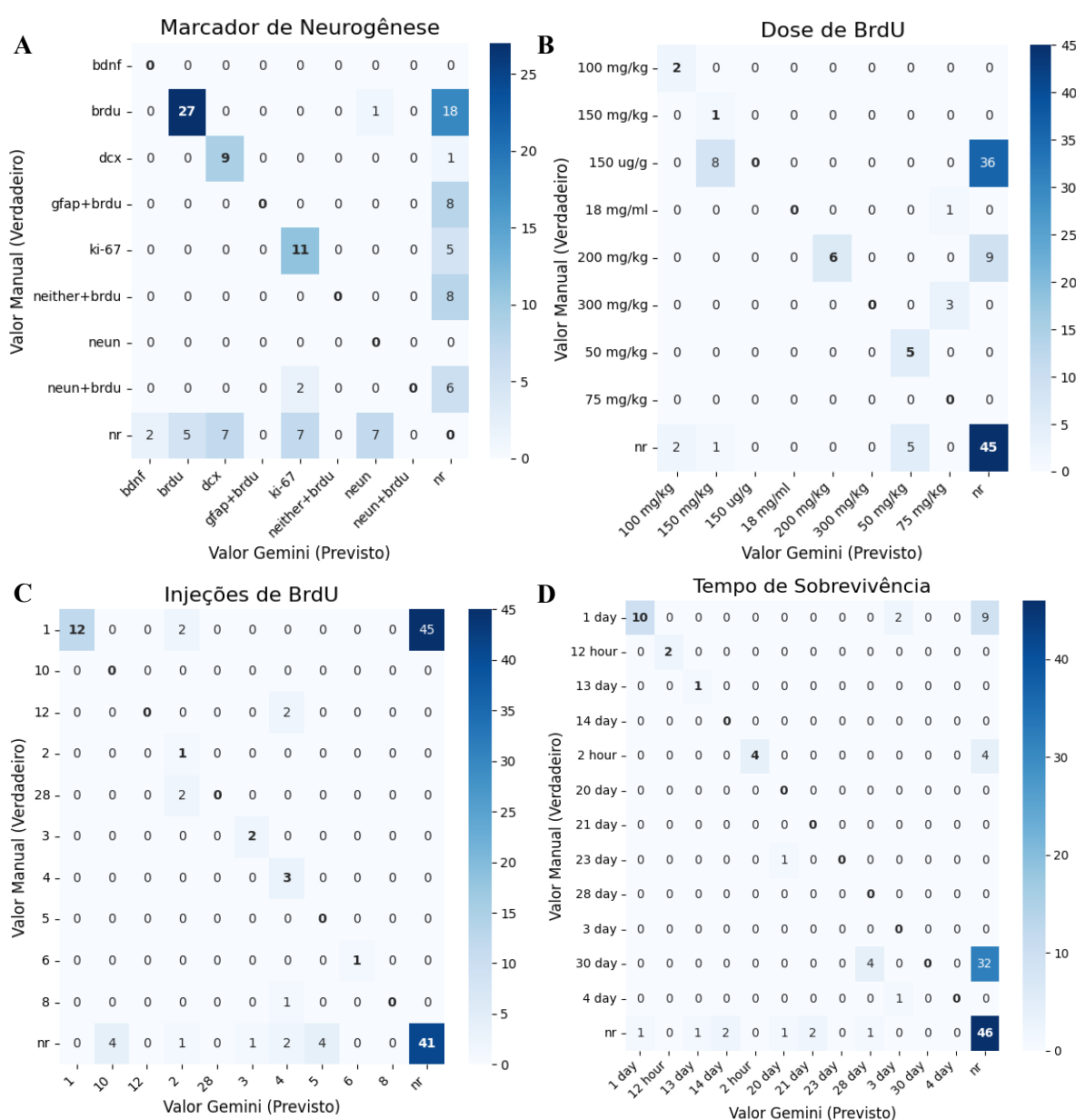


Figura 17. Matriz de confusão para marcador de neurogênese (a), número de injeções de BrdU (b), dose de BrdU (c) e tempo de sobrevivência (d) para os dados extraídos no Gemini. O eixo y (vertical) representa a

classificação do padrão-ouro ou referência de valor verdadeiro (Manual) e o eixo x (horizontal) a predição pelo modelo (GPT). A diagonal principal (negrito) indica acertos, e a coloração mais intensa das células indica a quantidade de *tokens* contados. NR: não reportado. *Elaborado pela autora (2025)*.

Para as análises de concordância das variáveis *multi-label*, em que o *score* Kappa ou as matrizes de confusão não são aplicáveis, foram construídos gráficos de halteres (*dumbbell plots*). Essa análise compara a frequência absoluta de cada categoria identificada pelo manualmente vs. ChatGPT e Gemini, evidenciando visualmente a magnitude das discrepâncias para cada teste comportamental ou protocolo de estresse.

Para os protocolos de estresse extraídos pelo ChatGPT e Gemini (**Figura 18**), observa-se que o ChatGPT alcançou uma concordância razoável ao considerarmos que os *tokens* não possuem uma disparidade numérica grande, variando de -3 a 2 discrepantes (**Figura 18a**). Em contrapartida, a única concordância total observada foi para o protocolo de estresse por contenção e imersão em água, representado pela sigla “wirs”, com a mesma frequência de *tokens*. O estresse por contenção (“srt”) e separação materna (“ms”) apresentaram extrações no Manual que estavam faltando no ChatGPT, ou seja, o modelo extraiu a menos. Em contrapartida, os protocolos de estresse crônico moderado (“cms”), estresse crônico moderado imprevisível (“cums”), estresse crônico imprevisível (“cus”), estresse precoce (“els”), modelo de depressão induzida por LPS (“lps”) e estresse por isolamento social (“sir”) foram extraídos a mais pelo modelo.

Na **Figura 18b** com os protocolos de estresse extraídos pelo Gemini, também se observa pouca discrepância ao considerar que os *tokens* não possuem uma disparidade numérica tão grande, com uma diferença máxima de 5 *tokens*. No entanto, o único protocolo de estresse que apresenta concordância total é o de estresse crônico moderado (“cms”). O restante dos *tokens* foram extraídos a menos (números positivos) ou foram extraídos a mais (números negativos) em relação à extração manual. Observa-se que o estresse por contenção (“srt”) é um valor faltante no Gemini, tendo sido extraído 2 vezes a mais no Manual. Adicionalmente, os protocolos de estresse crônico moderado imprevisível (“cums”), estresse crônico imprevisível (“cus”), estresse por contenção e imersão em água (“wirs”), oclusão da artéria cerebral média (“mcao”), separação materna (“ms”), labirinto aquático de Morris (“mwm”), rotenona (“rotenone”), estresse por isolamento social (“sir”) e modelo de depressão induzida por LPS (“lps”) são valores extraídos a mais pelo Gemini, i.e., estão ausentes na extração manual.

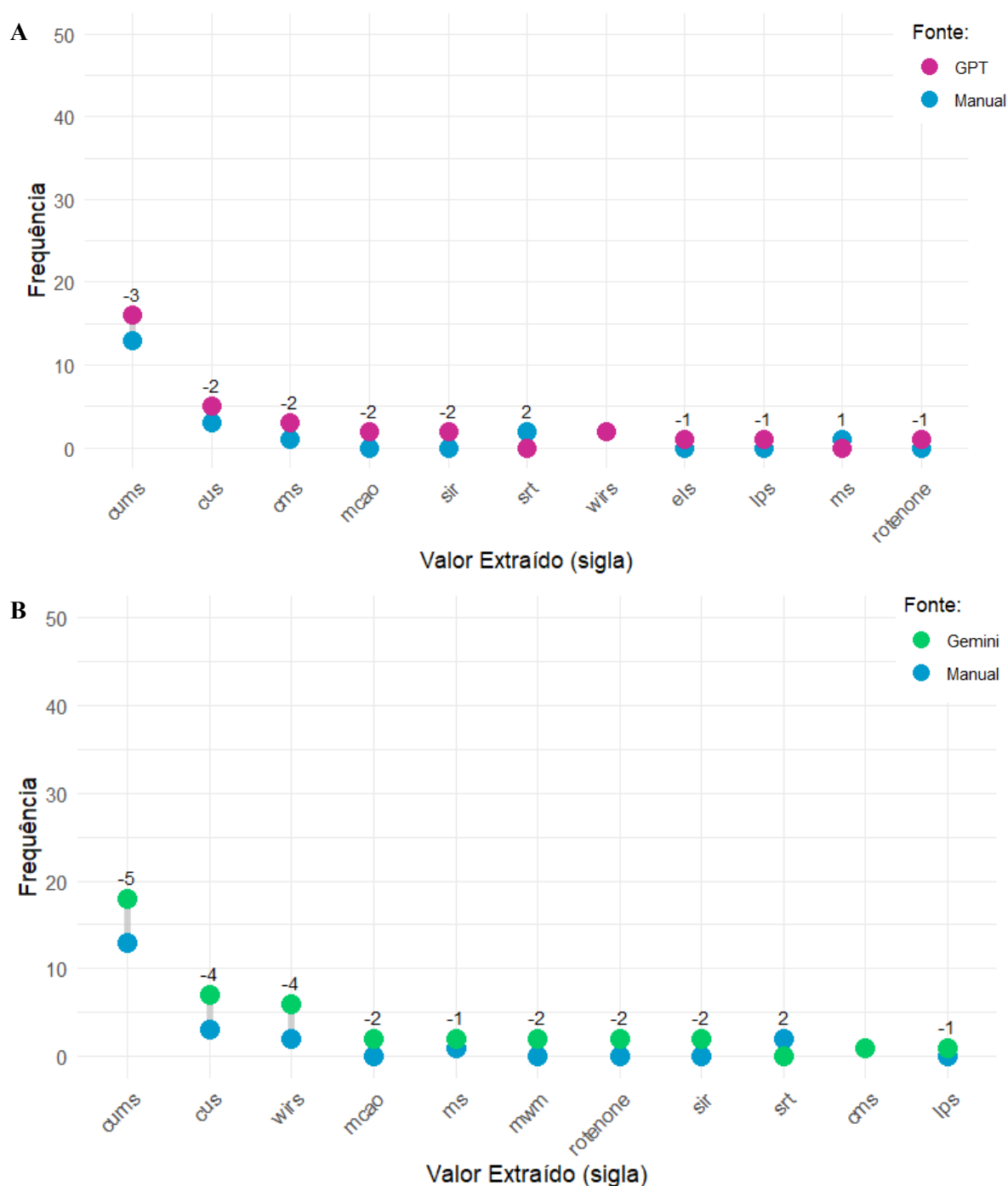
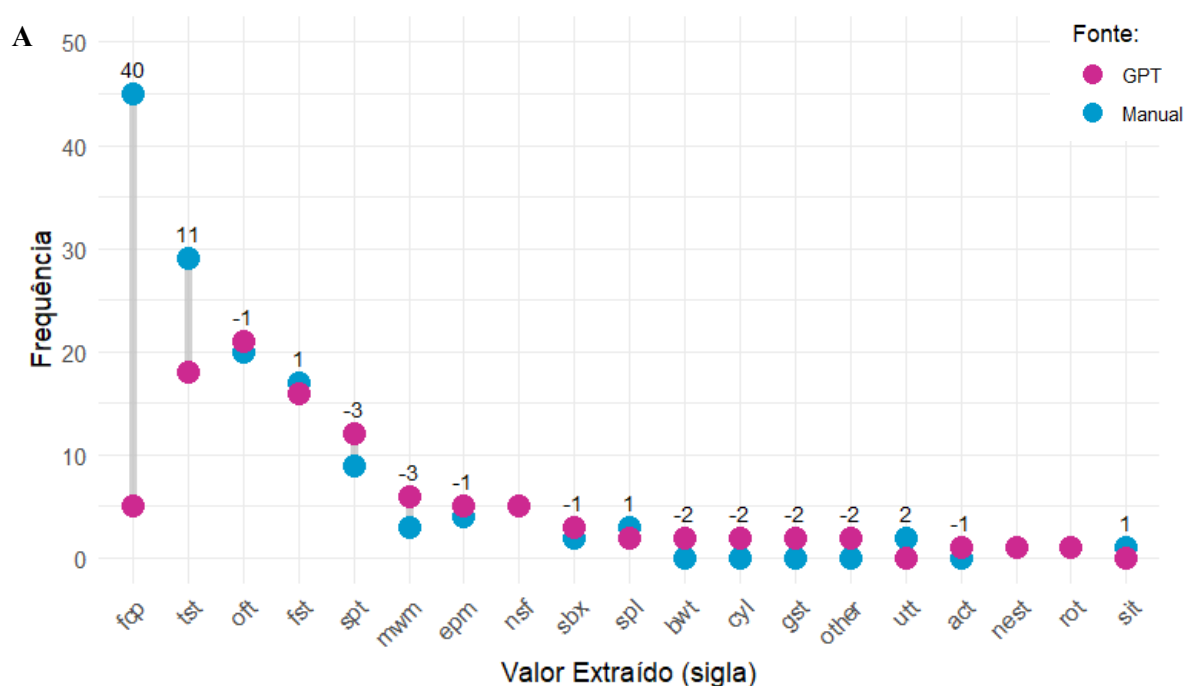


Figura 18. Concordância para protocolos de estresse extraídos manualmente vs. ChatGPT (a) e Gemini (b). O gráfico de halteres compara a contagem absoluta de cada termo extraído pelo Manual (azul), ChatGPT (rosa) ou Gemini (Verde). A distância vertical entre os pontos (conectados pela linha cinza) ilustra a divergência para cada tipo de protocolo de estresse. O número é igual a total de *tokens* divergentes, onde números negativos são *tokens* inventados pelos modelos e faltantes no Manual e números positivos são *tokens* omitidos pelos modelos e presentes no Manual. O eixo x (vertical) representa a frequência e o eixo y (horizontal) os *tokens* extraídos da variável de protocolo de estresse. CMS: Estresse Crônico Moderado; CUMS: Estresse Crônico Moderado Imprevisível; CUS: Estresse Crônico Imprevisível; ELS: Estresse Precoces; LPS: Modelo de Depressão Induzida por LPS; MS: Separação Maternal; SIR: Estresse por Isolamento Social; SRT: Estresse por Contenção; WIRS: Estresse por Contenção e Imersão em Água. *Elaborado pela autora (2025).*

Para os testes comportamentais extraídos pelo ChatGPT e Gemini (**Figura 19**), observa-se uma discrepância onde há *tokens* no Manual que não estão identificados pelo ChatGPT

(Figura 19a), sendo estes: teste comportamental de paradigma de condicionamento de medo (“fcp”), teste de suspensão pela cauda (“tst”), teste de nado forçado (“fst”), teste do borrito (“spl”), teste de interação social (“sit”) e teste de resposta inata ao som (“utt”). Além destes, dados referentes ao teste do campo aberto (“oft”), teste de preferência por sacarose (“spt”), labirinto aquático de Morris (“mwm”), teste do labirinto em cruz elevado (“epm”), teste da caixa de transição (“sbx”), teste de travessia na barra (“bwt”), teste do cilindro (“cyl”), teste de força de preensão (“gst”), teste de actímetro (“act”) e outros testes (“other”) foram extraídos pelo GPT, mas não estão presentes na planilha manual. Adicionalmente, os únicos *tokens* extraídos em concordância foram os testes de teste de construção de ninho (“nest”), teste de *rotarod* (“rot”) e teste de alimentação suprimida pela novidade (“nsf”).

O Gemini também demonstrou dificuldade de identificar corretamente alguns testes comportamentais (Figura 19b). Isso pode ser observado para o teste de paradigma de condicionamento de medo (“fcp”), teste de suspensão pela cauda (“tst”) e teste de resposta inata ao som (“utt”) que foram perdidos pelo Gemini. Em contrapartida, os *tokens* para o teste de nado forçado (“fst”), labirinto aquático de Morris (“mwm”), teste de preferência por sacarose (“spt”), labirinto em cruz elevado (“epm”), alimentação suprimida pela novidade (“nsf”), teste da caixa de transição (“sbx”), teste de interação social (“sit”), teste de travessia na barra (“bwt”), teste do cilindro (“cyl”), teste de força de preensão (“gst”) e *rotarod* (“rot”) foram extraídos pelo Gemini, mas não manualmente. Adicionalmente, houve concordância nas extrações para o teste de campo aberto (“oft”), teste do borrito (“spl”) e teste de construção de ninho (“nest”).



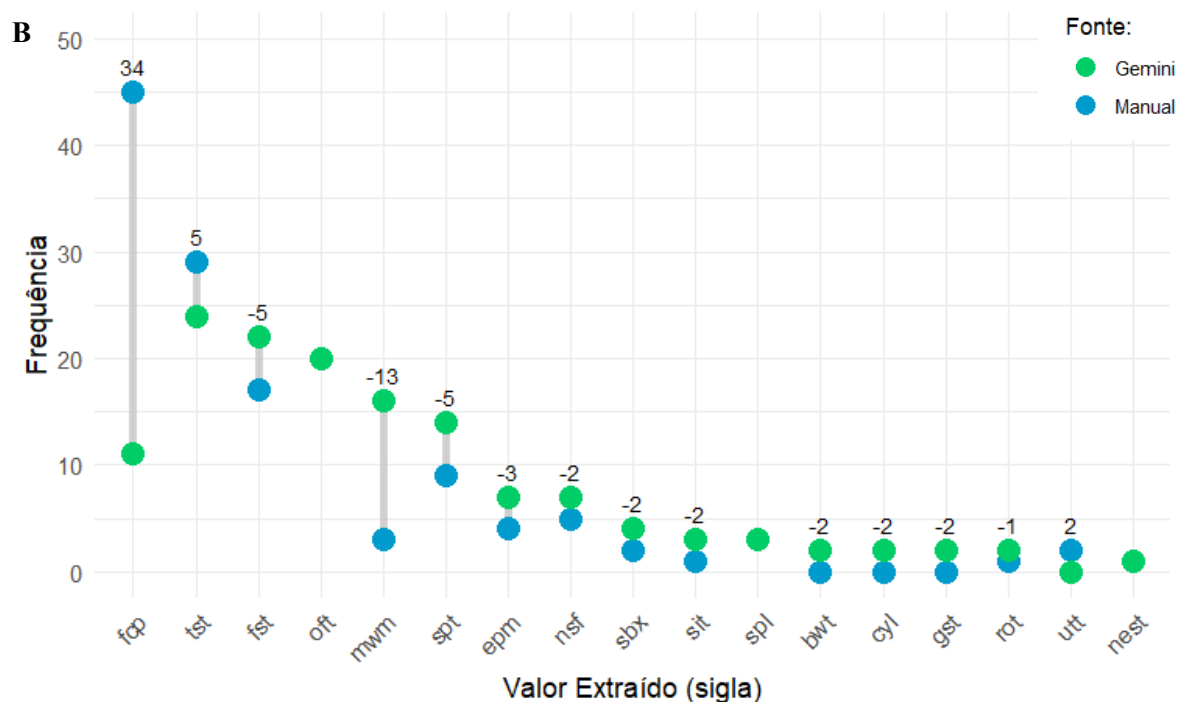


Figura 19. Concordância para testes comportamentais extraídos manualmente vs. ChatGPT (a) e Gemini (b). O gráfico de halteres compara a contagem absoluta de cada termo extraído pelo Manual (azul), ChatGPT (rosa) ou Gemini (Verde). A distância vertical entre os pontos (conectados pela linha cinza) ilustra a divergência para cada tipo de teste comportamental. O número é igual a total de *tokens* divergentes, onde números negativos são *tokens* inventados pelos modelos e faltantes no Manual e números positivos são *tokens* omitidos pelos modelos e presentes no Manual. O eixo x (vertical) representa a frequência e o eixo y (horizontal) os *tokens* extraídos da variável de teste comportamental. ACT: Teste de Actímetro; BWT: Teste de Travessia na Barra; CYL: Teste do Cilindro; EPM: Teste de Labirinto em Cruz Elevado; FCP: Teste de Paradigma de Condicionamento do Medo; FST: Teste de Nado Forçado; GST: Teste de Força de Preensão; MWM: Teste do Labirinto Aquático de Morris; NEST: Teste de Construção de Ninho; NSF: Teste de Alimentação Suprimida pela Novidade; OFT: Teste de Campo Aberto; ROT: Teste do Rotarod; SBX: Teste da Caixa de Transição; SIT: Teste de Interação Social; SPL: Teste do Borrifo; SPT: Teste de Preferência por Sacarose; TST: Teste de Suspensão pela Cauda; UTT: Teste de Resposta Inata ao Som. *Elaborado pela autora (2025).*

6 DISCUSSÃO

6.1 Extração de dados automatizada para revisões sistemáticas na farmacologia

O principal empecilho identificado na literatura para o avanço das revisões sistemáticas é o tempo exaustivo dedicado à extração de dados (Jonnalagadda et al., 2015; Elbadawi et al., 2024). Os resultados do presente estudo demonstram que a aplicação de LLMs (ChatGPT e Gemini) atua sobre esse gargalo, como pôde ser demonstrado pela disparidade entre o tempo despendido na extração manual (~26 horas de trabalho ativo) e o processamento pelos modelos de linguagem (~10 segundos por ciclo de extração), representando uma mudança de paradigma na viabilidade das revisões sistemáticas. No contexto da pesquisa pré-clínica, essa economia de tempo é particularmente valiosa quando consideramos que a extração manual não é apenas lenta, mas vulnerável à fadiga cognitiva, um fator que aumenta a probabilidade de erro humano em tarefas repetitivas e exaustivas (Bannach-Brown et al., 2019). Diferente do avaliador

humano, cuja acurácia pode oscilar conforme o volume de artigos processados, os LLMs mantêm uma velocidade de processamento constante, permitindo que o ganho de eficiência seja escalonado sem o ônus do desgaste intelectual.

A aceleração do tempo também mitiga o risco de obsolescência da revisão sistemática, como apontado por Shojania et al. (2007). A automação permite vislumbrar um cenário de “revisões vivas” (*living systematic reviews*) onde, neste caso, a síntese de evidências poderia ser atualizada em tempo real conforme novos estudos sobre neurogênese e antidepressivos são indexados, rompendo o gargalo histórico da síntese manual exaustiva. Ao projetarmos essa economia para bibliotecas maiores — como a de Bolzan (2022) com dezenas de publicações — o ganho de eficiência deixa de ser um mero facilitador e se torna o um meio viável de manter a evidência científica sincronizada com a produção acadêmica abundante. Ao considerarmos que o volume de publicações anuais na área biomédica ultrapassa um milhão de artigos, a capacidade dessas ferramentas de processar bibliotecas inteiras em um único ciclo, como é o caso do Gemini por sua janela de contexto massica, validaria o conceito da “publômica” (Ineichen, Rosso & Macleod, 2023).

Nesse contexto, os resultados de nosso estudo confirmam que a IA não é apenas capaz de extrair os dados, mas também torna viável o manejo de grandes volumes de informação, que é cognitivamente cansativo e, em alguns casos, impossível para um revisor humano. Portanto, a problemática central não reside na capacidade da IA em mitigar o gargalo operacional, mas no compromisso entre velocidade e acurácia, visto que a agilidade obtida impõe uma necessidade de supervisão humana rigorosa diante das margens de erro inerentes ao processamento automatizado (Gartlehner et al., 2024; Alshami et al., 2023; Messeri & Crockett, 2024). Entretanto, a questão de se a inteligência artificial é capaz de resolver esse gargalo exige uma análise olhando para outra nuance: a eficiência temporal realmente é incontestável (Bannach-Brown et al., 2019; Alshami et al., 2023; Clark et al., 2020), mas os resultados mostram que o gargalo muda de natureza. Ele deixa de ser um problema de “leitura e extração” para se tornar um problema de “curadoria”. Embora o processamento seja instantâneo, a necessidade de um processamento de dados via código sugere que o tempo poupado na extração ainda é reinvestido na manipulação dos dados para que estes consigam ser analisados com a devida fidelidade.

No presente estudo, embora o desempenho tenha sido alto para dados qualitativos (F_1 -score > 80%), a precisão para metadados quantitativos necessita de uma camada maior de

tratamento algorítmico pós-extração para garantir a integridade da meta-análise subsequente. Assim, é fundamental diferenciar o tempo de execução do tempo de fluxo de trabalho. Os resultados apontam para um modelo híbrido de extração de dados, confirmando o que autores como Marshall & Wallace (2019) defendem: um sistema onde a IA executa o trabalho inicial pesado de extração e o pesquisador atua na etapa de conciliação dos dados. Assim, a supervisão humana garantiria que “estudos perdidos” (observados como falhas estruturais) sejam recuperados e que os dados sejam padronizados e não se percam em decorrência de variações terminológicas que são consequência da variabilidade estocástica dos LLMs. Portanto, a supervisão humana deixa de ser uma tarefa de mineração para se tornar uma função de auditoria. O tempo despendido tanto no tratamento e padronização dos dados quanto na conferência de pares gerados pelo algoritmo de alinhamento é qualitativamente distinto e menos exaustivo do que a busca manual de dados em textos não estruturados, permitindo que o pesquisador foque seu esforço intelectual na interpretação crítica dos resultados e não na coleta repetitiva de variáveis.

6.2 Desenvolvimento do protocolo para extração de dados com modelos de IA

Os modelos ChatGPT e Gemini foram selecionados para desenvolvimento da ferramenta por apresentarem a maior parte dos requisitos necessários para cumprir os objetivos deste trabalho, tais como interface amigável, capacidade de leitura e acesso gratuito. As outras LLMs avaliadas apresentaram qualidades benéficas, porém, falharam em requisitos importantes, tais como acesso pago ou necessidade de infraestrutura no laboratório. Por exemplo, embora o modelo Claude 4 Opus apresentasse elevada capacidade de leitura, o custo de API era significativamente maior (\$75/1M *tokens* de saída). Neste sentido, o custo financeiro poderia inviabilizar a escalabilidade do protocolo para grandes bibliotecas e restringir seu acesso. Similarmente, o Llama 3 requereria uma infraestrutura de *hardware* local (e.g., GPU com alta VRAM), dificultando o objetivo de disponibilizar uma ferramenta acessível aos pesquisadores. O ChatGPT, além de ser o modelo mais “popular”, também permite fácil acesso à sua interface de *chat*. O Gemini 2.5 Pro se destacou ao apresentar a maior janela de contexto entre os modelos (>2M *tokens*), uma característica técnica fundamental para a leitura integral de múltiplos artigos científicos simultaneamente sem perda de coerência. Portanto, a escolha dos modelos se pautou tanto na capacidade de leitura massiva de artigos (Gemini) quanto na acessibilidade (GPT).

A construção do protocolo para extração de dados exigiu uma engenharia de *prompt* iterativa, fundamentada em técnicas de Processamento de Linguagem Natural (NLP) para

mitigar a natureza estocástica dos modelos. A atribuição de uma *persona* (revisor sistemático) e a estruturação do comando sob a lógica PICO foram táticas para restringir o espaço probabilístico de resposta da IA. Como os LLMs operam prevendo o próximo *token* mais provável, a inclusão de diretrizes proibitivas e a instrução explícita para o uso do rótulo “nr” (não reportado) serviram para forçar o modelo a admitir a ausência de informação em vez de preenchê-la com dados estatisticamente plausíveis, porém falsos. Esse refinamento buscou uma padronização de extração automatizada para a área farmacológica, almejando um equilíbrio entre a liberdade semântica do modelo e o rigor metodológico da síntese de evidências. Mesmo com estes cuidados, como discutido adiante, houve circunstâncias em que a aplicação do *prompt* retornou informações falsas, ou “alucinações”.

A demonstração da aplicabilidade do protocolo no ChatGPT serviu como uma prova de conceito fundamental para validar se a estrutura PICO, aliada a diretrizes de controle de alucinação, era capaz de navegar pela densidade de artigos da farmacologia. Os resultados revelaram que a ferramenta é altamente aplicável para a extração rápida, convertendo a leitura técnica em dados estruturados em segundos. Contudo, a aplicabilidade prática enfrentou barreiras extrínsecas ao modelo: as restrições de *upload* e leitura do plano gratuito. Este achado é metodologicamente relevante, pois indica que a eficiência de uma revisão sistemática automatizada não depende apenas da acurácia dos LLMs, mas da logística de fluxo permitida pela interface. Seria necessário testar se a eficiência do *prompt* aumentaria em um plano pago do ChatGPT.

Em termos de transferibilidade, a aplicação do protocolo para o modelo Gemini 2.5 Pro sem a necessidade de uma nova engenharia de *prompt* confirmou a robustez e a eficiência do comando desenvolvido. A transferibilidade foi considerada bem-sucedida não apenas pela execução do comando, mas pelo incremento notável na qualidade da extração. Diferente do ChatGPT, o Gemini operou sob uma arquitetura de janela de contexto massiva, processando a amostra com maior geração de linhas (estudos) para os dados brutos. Dessa forma, a superioridade do Gemini em evitar omissões evidencia que modelos com suporte robusto de processamento são menos suscetíveis à fragmentação da informação. O achado alinha-se às evidências de que o desempenho em tarefas de raciocínio complexo escala positivamente com o tamanho do modelo (Kaplan et al., 2020; Wei et al., 2022), permitindo que a ferramenta sustente a coerência necessária para a extração detalhada de diversos braços experimentais em uma revisão sistemática. Portanto, a transferibilidade validada aqui indica que o pesquisador pode transitar entre diferentes plataformas de IA buscando o melhor custo-benefício ou

eficiência de processamento, sem comprometer a padronização metodológica do protocolo original, com ciência de que tipo de plano pode afetar a qualidade dos dados extraídos. Afinal, independentemente do modelo aplicado, foram observados defeitos estruturais nos dados tabulados pelos LLMs e erros de conteúdo contido nestes dados.

6.3 Validade estrutural dos dados extraídos pelos modelos de IA

A observação prévia das planilhas de dados criadas pelo protocolo de extração de dados usando os modelos ChatGPT e Gemini 2.5 Pro revelou erros de tabulação que comprometeriam a análise futura destes dados. Por exemplo, observou-se uma tendência, especialmente no ChatGPT, em aglutinar múltiplos experimentos em uma única célula com o uso de separadores, resultando em apenas 35 linhas iniciais. Em contrapartida, o modelo Gemini apresentou maior aderência à regra de estruturação tabular definida no *prompt*, aproximando-se um pouco mais do volume de estudos obtido na extração manual (58 linhas vs. 74 linhas). No entanto, a contagem de linhas isolada não foi métrica suficiente para avaliar a sensibilidade da IA, tendo sido necessário um protocolo que incluísse, obrigatoriamente, etapas de desagregação (via função *explode* em Python/Pandas) e algoritmos de atribuição linear (*best-fit*). Estes fatos reforçam a necessidade da constante supervisão humana, não apenas para corrigir erros, mas para estruturar a lógica biológica que os modelos atualmente não possuem. Além do mais, o tratamento dos dados pós-extração foi realizado aplicando *scripts* em linguagens R e Python no *software* Rstudio. Embora os *scripts* tenham sido formulados com pacotes de domínio público e rodados em um *software* considerado intuitivo, é necessário ter alguma familiaridade com programação para o trabalho. Desta forma, a versão atual do protocolo de extração de dados com os modelos ChatGPT e Gemini 2.5 Pro requer que os usuários tenham alguma experiência com programação para garantir a integridade dos dados extraídos de modo automatizado.

Em suma, a eficácia do algoritmo de alinhamento *best-fit* foi o que permitiu que as métricas de desempenho fossem calculadas. Sem o realinhamento das linhas — que corrigiu inversões em 18 casos para o GPT e 31 para o Gemini — o desempenho dos modelos seria artificialmente subestimado devido a questões puramente estruturais de ordem de saída. Apesar disso, os resultados mostram que a estrutura dos dados extraídos pelas IAs não é inteiramente confiável, mesmo quando obtidos com versões pagas. Isso demonstra que a validação de futuras ferramentas de IA na ciência deve considerar a arquitetura dos dados gerados, validando o protocolo híbrido (IA + *scripts*) como o padrão necessário para a automação confiável de revisões sistemáticas pré-clínicas.

6.4 Validade do conteúdo dos dados extraídos pelos modelos de IA

Um debate central na literatura de IA científica é se modelos generalistas, como o ChatGPT e Gemini, podem competir com modelos especializados no domínio biomédico, como o PubMedBERT (Gu et al., 2021; Gutierrez et al., 2022). Os resultados apresentados neste trabalho trazem uma contribuição para esse debate. Por mais que o Gemini e o ChatGPT tenham obtido concordância “quase perfeita” em espécie, linhagem e antidepressivo, sugerindo que a vasta base de dados de treinamento desse modelo já engloba o vocabulário técnico da farmacologia, as limitações surgiram na inferência de protocolos complexos. O fato de o GPT ter atingido apenas 36% de F_1 -score no número de injeções de antidepressivo revela que, enquanto o vocabulário é bem compreendido, a lógica experimental ainda é um desafio para modelos generalistas. A penalização sofrida no score Kappa para variáveis numéricas como “duração do tratamento” e “número de injeções” reflete o desafio da máquina em processar a aritmética experimental. Enquanto o avaliador humano interpreta que um protocolo de injeções de “P3 a P15” totaliza 12 dias, a IA muitas vezes extrai o termo literal sem realizar a conversão temporal necessária. Sendo assim, estratégias híbridas são novamente preferíveis, onde a combinação de extração semântica por IA e a normalização quantitativa por regras determinísticas se mostram mais adequadas para revisões sistemáticas.

Ao comparar o desempenho de ambos os modelos, o Gemini 2.5 Pro demonstrou uma superioridade métrica em relação ao ChatGPT na maioria das variáveis PICO. Enquanto o ChatGPT apresentou uma tendência maior à fragmentação e à omissão, apresentando dificuldades em ultrapassar o limiar de 50% para doses e durações, o Gemini exibiu uma sensibilidade significativamente mais alta, refletida em F_1 -scores que ultrapassaram 80% em variáveis críticas de intervenção, como o tipo de antidepressivo e a dose. O Gemini também se mostrou mais resiliente na extração de variáveis de rótulos múltiplos (*multi-label*), como testes comportamentais e protocolos de estresse, que exigem uma varredura exaustiva ao longo do artigo. Tais achados corroboram a literatura sobre habilidades emergentes (Wei et al., 2022), sugerindo que modelos de maior escala lidam melhor com complexidade textual, inerente a múltiplos braços experimentais. Além disso, a janela de contexto massiva do Gemini permitiu correlacionar informações dispersas sem sofrer do fenômeno “*lost in the middle*” (Liu et al., 2024), que descreve a “perda da atenção” em modelos de linguagem quando a informação relevante está situada no meio de contextos extensos, uma limitação evidente no comportamento do ChatGPT. Em contrapartida, a persistência do rótulo “nr” em ambos os modelos para dados existentes nos artigos sinaliza uma barreira na interpretação de

informações. Isso indica que, quando a informação está presente apenas em formato visual (figuras ou tabelas), a capacidade de extração da IA é comprometida. Essa limitação é particularmente relevante na literatura farmacológica, onde parâmetros experimentais críticos são frequentemente apresentados de forma não textual, dificultando a extração via LLMs. Isso reforça que, apesar do avanço em janelas de contexto (Kaplan et al., 2020), a interpretação de dados puramente visuais e tabelas complexas ainda representa um gargalo para a automação.

O Gemini, embora métrica e tecnicamente mais preciso, demonstrou uma tendência ao excesso na extração de desfechos, capturando a totalidade dos testes comportamentais relatados nos artigos. Em contrapartida, a extração manual operou sob um filtro de relevância, selecionando apenas testes diretamente vinculados à atividade antidepressiva. A capacidade superior do Gemini em mapear variáveis *multi-label* de alta complexidade reflete o desenvolvimento de habilidades emergentes que surgem apenas em modelos de grande escala (WEI et al., 2022). Esse fenômeno explica por que o Gemini obteve F_1 -scores elevados, mas apresentou um volume maior de falsos positivos qualitativos em variáveis *multi-label*, uma vez que o modelo priorizou a recuperação integral em detrimento da filtragem seletiva que um avaliador humano aplicaria. As análises diagnósticas das matrizes de confusão revelam que o viés predominante em ambos os modelos não é a alucinação, ou invenção deliberada de dados, mas sim a falha de recuperação e a declaração dessa ausência da informação, gerando um falso negativo de recuperação que gera um falso positivo pela afirmação do *token* “nr”.

A variabilidade observada entre os modelos é corroborada por evidências recentes que comparam a eficácia do ChatGPT, DeepSeek e Gemini em tarefas de revisões sistemáticas e revisões guarda-chuva (*umbrella reviews*). Segundo Emami e Shirani (2025), embora modelos como o ChatGPT e o DeepSeek possam apresentar taxas de precisão elevadas na extração de dados brutos (>88%), o Gemini destaca-se por uma curva de aprimoramento significativo em sua performance ao longo do tempo. No entanto, o estudo ressalta que a confiabilidade em tarefas críticas, como a triagem de texto completo e a avaliação de risco de viés, permanece um desafio comum a todos os LLMs, o que reforça a necessidade de protocolos de validação robustos, como o aplicado no presente estudo, para mitigar o risco de omissões e garantir a integridade da síntese de evidências (Emami & Shirani, 2025).

Os erros observados seguiram padrões de proximidade semântica, onde os modelos tendem a confundir categorias dentro de um mesmo domínio, como linhagens de ratos com características fenotípicas similares. Tal fenômeno é amplamente documentado como uma

limitação na distinção de entidades granulares dentro de espaços latentes próximos (Borji, 2023). Além disso, os resultados confirmam que o vocabulário qualitativo é mais acessível à IA do que a precisão quantitativa, uma vez que LLMs priorizam a verossimilhança linguística em detrimento do rigor matemático (Geva, Gupta & Berant, 2020). O erro sistemático em classificar doses como “nr” sugere uma falha na identificação de padrões numéricos associados a unidades de medida quando estas não aparecem imediatamente adjacentes ao nome do fármaco, evidenciando a dificuldade dos modelos em manter dependências de longo prazo em extrações estruturadas (Thawani et al., 2021). Além do mais, a insistência no erro “nr” ao invés de retornar o número de injeções reforça a hipótese de que os LLMs falham em converter expressões linguísticas (e.g. *single*, *one*) em valores numéricos simbólicos, um obstáculo conhecido como a “barreira dos números” em modelos de linguagem (Wallace et al., 2019).

A dispersão dos dados fora da diagonal principal nas matrizes de duração do tratamento e injeções revela que os modelos frequentemente confundem protocolos temporais próximos (como 12 vs. 14 dias). Similarmente, as matrizes de tempo de sobrevivência e injeções de BrdU reforçam a dificuldade dos modelos em identificar detalhes de protocolos. O padrão de erro indica que protocolos experimentais de longa duração ou descritos de forma mais diluída no texto são menos visíveis para os modelos do que protocolos agudos e imediatos, o que pode ser explicado pela degradação da atenção em passagens textuais densas (Saparov & He, 2022). Finalmente, as matrizes de confusão validam que, embora o Gemini seja tecnicamente mais sensível que o GPT, ambos operam sob um regime de prudência excessiva. Isso resulta em uma base de dados automatizada com alta precisão (poucos dados errados inseridos), mas baixa sensibilidade (muitos dados corretos ignorados), refletindo a calibração conservadora dos modelos modernos para evitar alucinações (Kadavath et al., 2022; Zhao et al., 2021).

A aplicação das métricas de desempenho evidenciou o porquê de o F_1 -score ser um indicador mais fidedigno da utilidade prática da ferramenta do que o índice Kappa em certos contextos. Para variáveis *multi-label*, o F_1 -score permitiu capturar os “acertos parciais” (e.g., quando a IA identificou corretamente um “fst”, mas omitiu um “tst”). Nesses casos, a análise por meio de gráficos de halteres confirmou que a magnitude da divergência é menor para testes paradigmáticos da depressão, indicando que a IA possui uma base de conhecimento mais sólida para termos universais da farmacologia. Por outro lado, a severa penalização do Kappa em variáveis binárias como o sexo reforça que, para dados de alta criticidade e baixa variabilidade, qualquer erro de inferência da máquina é estatisticamente punido, exigindo que o pesquisador aplique uma camada de validação humana redobrada sobre esses parâmetros antes de prosseguir

para a síntese quantitativa. Modelos especializados podem ser mais precisos em extrair relações de causa e efeito, mas a acessibilidade e a fluidez dos LLMs via interface de *chat* democratizam a automação para pesquisadores que não dominam linguagens de programação. Somado a isso, a popularidade dos LLMs é justificada também por estudos recentes sugerindo que a escala massiva de parâmetros de modelos generalistas atuais pode mitigar a necessidade de um pré-treinamento exclusivo em bases científicas, embora ainda apresentem desafios de fidelidade em tarefas de extração complexas (Nori et al., 2023).

Em resumo, além de defeitos estruturais corrigidos pelo trabalho pós-extração, a versão atual do protocolo de extração de dados com os modelos ChatGPT e Gemini 2.5 Pro apresentado neste trabalho demonstra que os modelos de IA cometeram erros de “interpretação” que comprometem a validade dos dados extraídos automaticamente sem a curadoria humana. Os resultados sugerem que as informações obtidas a partir dos dados extraídos pelas IAs, mesmo após a manipulação dos dados e mesmo em versões pagas, não são inteiramente confiáveis. Isso reforça que a validação de ferramentas de IA na ciência não deve se limitar à verificação estrutural e de arquitetura dos dados gerados, mas deve considerar a validade do conteúdo extraído como o padrão necessário para a automação confiável de revisões sistemáticas pré-clínicas.

6.5 Aplicação de modelos de IA em uma revisão sistemática de neurofarmacologia

A metodologia de revisões sistemáticas exige, tradicionalmente, dois revisores independentes para garantir a fidedignidade dos dados. A investigação sobre a viabilidade de substituir um desses revisores por uma inteligência artificial revelou que os modelos testados demonstraram uma tendência maior à omissão (falsos negativos) do que à alucinação (falsos positivos), padrão recorrente em modelos submetidos a protocolos rigorosos de alinhamento de segurança, que priorizam a abstenção à geração de conteúdo errôneo (Tang et al., 2023). No caso do artigo ID 52, por exemplo, onde foram identificados 44 estudos manualmente e o GPT recuperou apenas 2, fica evidente que a IA ainda falha em detectar a complexidade de delineamentos experimentais fatoriais. O uso frequente do rótulo “nr” em locais onde o dado estava presente indica um “viés de prudência” induzido pelo *prompt* de restrição (Kadavath et al., 2022). Essa dificuldade de recuperação integral também parece estar intrinsecamente ligada à heterogeneidade e à baixa qualidade no relato da pesquisa biomédica. Como apontado por Macleod et al. (2014), a ciência pré-clínica sofre de uma crise decorrente de descrições metodológicas incompletas e falta de padronização terminológica. Para uma IA, que opera por padrões probabilísticos, a ausência de uma estrutura de relato uniforme, como a proposta pelas

diretrizes ARRIVE, atua como um ruído que impede a desambiguação de dados complexos. A heterogeneidade na apresentação de resultados, onde um mesmo desfecho pode estar em um gráfico, em uma tabela ou diluído no texto, impõe um limite à eficácia dos LLMs, que ainda não possuem o raciocínio sistêmico para integrar essas fontes dispersas de forma confiável (Macleod et al., 2015).

Enquanto o domínio clínico consolidou, ao longo de décadas, diretrizes de relato rigorosas e amplamente adotadas, como o CONSORT para ensaios clínicos randomizados (Moher et al., 2010), a literatura biomédica ainda padece de uma fragmentação. Na pesquisa clínica, os manuscritos seguem uma estrutura quase tabular e previsível, o que facilita a ancoragem semântica da IA. Em contrapartida, na farmacologia pré-clínica, variáveis biológicas e metodológicas fundamentais são frequentemente omitidas ou descritas de forma assistemática, sem padronização (Percie du Sert et al., 2020). Para um modelo de linguagem, essa falta de padronização, conhecida como ruído de relato, transforma a extração em um desafio de inferência probabilística em vez de uma recuperação direta de fatos. Quando um pesquisador humano lê um artigo pré-clínico, ele realiza automaticamente uma “limpeza” dos dados, interpretando variações terminológicas e localizando informações dispersas em legendas ou materiais suplementares. A IA, por outro lado, ao se deparar com a dispersão de dados em textos mal estruturados, perde os pontos de referência estatística. Como resultado, o modelo tende a adotar o regime de prudência excessiva diante da incerteza gerada pelo relato assistemático (Kadavath et al., 2022), o que confirma que a baixa qualidade da redação científica original é um dos maiores obstáculos para a automação da síntese de evidências (Macleod et al., 2015). Esse comportamento indica que a IA, no presente estágio, não pode atuar como um “segundo revisor” autônomo, pois sua taxa de omissão comprometeria a sensibilidade da revisão. Por outro lado, a precisão alcançada em variáveis como espécie (F_1 -score de 98%) sugere que a IA pode atuar como um “auxiliar” de alta performance em uma etapa de triagem e extração de dados semiautomatizada. Nesse modelo de “humano-no-ciclo” (*human-in-the-loop*), o sistema computacional reduz drasticamente a carga de trabalho braçal, mas o julgamento final, a interpretação contextual e a auditoria permanecem sob responsabilidade humana para garantir a validade científica e mitigar os efeitos da heterogeneidade (Marshall & Wallace, 2019).

Os índices de concordância medidos pelo Kappa de Cohen corroboram a percepção de que a confiabilidade da IA é dependente da clareza terminológica do dado. A concordância foi classificada como quase perfeita para variáveis como espécie (100%), marcador de neurogênese

(82% no GPT e 90% no Gemini) e linhagem (76% no GPT e 98% no Gemini). No entanto, os resultados sugerem que a concordância em variáveis binárias aparentemente simples, como o sexo dos animais (11% no GPT e 28% no Gemini), expõe uma falha crítica de inferência. Foi observado que ambos os modelos de IA frequentemente falhou em detectar quando o estudo utilizava ambos os sexos (“both”), tendendo a assumir o “macho” como padrão ou a omitir a informação (“nr”). Um motivo para isso pode ser um viés de treinamento dos LLM e pelo persistente viés de gênero na pesquisa biomédica pré-clínica, onde o uso de animais machos predomina amplamente na literatura para evitar a suposta variabilidade induzida pelo ciclo estral das fêmeas (Beery & Zucker, 2011; Woitowich et al., 2020). Considerando que os LLMs operam através de cálculos probabilísticos sobre vastos conjuntos de dados (Bender et al., 2021; Kaplan et al., 2020; Vaswani et al., 2017; Zhao et al., 2023), é provável que, ao encontrar ambiguidades ou dificuldades na localização explícita do termo no texto, os modelos tenham inferido o sexo masculino por este ser o *token* estatisticamente mais frequente no domínio da neuropsicofarmacologia. Em vez de uma extração neutra, o que se observa é uma reprodução automatizada do viés sistêmico da área, onde a máquina preenche a lacuna de informação com o cenário mais recorrente em seu treinamento.

As alucinações de variáveis qualitativas, como a identificação de testes comportamentais que não constavam no protocolo original do artigo, podem sugerir um efeito de “preenchimento por associação”. Por exemplo, ao identificar o uso de fluoxetina e protocolos de estresse, o modelo pode ter associado probabilisticamente a presença do teste de nado forçado, mesmo que este não tenha sido executado (Holtzman et al., 2021), agindo como o “papagaio estocástico” descrito por Bender et al. (2021) que prevê o desfecho mais provável em estudos de depressão. Ao avaliar a qualidade da extração por meio das métricas de F_1 -score, precisão e sensibilidade (revocação), observa-se uma hierarquia de desempenho clara entre variáveis qualitativas e quantitativas. Ambos os modelos atingiram um desempenho classificado como “quase perfeito” para variáveis categóricas simples, como espécie, alcançando F_1 -scores entre 98% e 100%, com precisão e sensibilidade máximas. Isso sugere que a identificação taxonômica é uma tarefa de Reconhecimento de Entidades Nomeadas (*Named Entity Recognition*, NER) já consolidada para LLMs (Li et al., 2018).

Somado a isto, observa-se um declínio acentuado na identificação de parâmetros farmacológicos que exigem precisão numérica e inferência lógica. O modelo GPT, por exemplo, demonstrou fragilidade extrema ao capturar o número de injeções de antidepressivo (F_1 -score de 36%) e a duração do tratamento (F_1 -score de 41%). A baixa sensibilidade

observada nestes campos específicos indica que o modelo frequentemente falha em localizar doses e frequências quando estas não estão explicitamente adjacentes ao nome do fármaco, optando por rotular a informação como “nr”. Por exemplo, é comum observar nos artigos incluídos a descrição “*animals received a dose of...*” em vez de “*animals received an injection of...*”. Ao se deparar com tal situação, o ser humano facilmente interpreta o que é relatado e extrai a informação corretamente. Particularmente na teoria neurogênica, a confusão entre marcadores (BrdU, Ki-67, DCX) e seus respectivos protocolos de sobrevivência revelou um ponto de fragilidade no raciocínio de cadeia (*chain-of-thought*), onde os modelos operam como “raciocinadores gananciosos” que perdem a coesão em protocolos complexos (Saparov & He, 2022). O Gemini, embora mais robusto, apresentou alucinações em doses específicas (e.g., convertendo 150 ug/g para 150 mg/kg), uma mudança de unidade que, embora matematicamente equivalente, demonstra que o modelo “interpreta” com base em seu pré-treinamento. Essa tendência de reescrever dados para unidades mais frequentes na literatura (como mg/kg) reflete o viés de popularidade, onde a IA ignora informações raras ou unidades atípicas em favor de padrões estatisticamente dominantes em sua base de treinamento (Kandpal et al., 2023). Essa dependência da frequência estatística dos termos também se reflete na dificuldade em distinguir linhagens genéticas específicas em comparação à facilidade com linhagens comerciais, i.e., quanto mais raro o modelo animal na literatura geral, maior a probabilidade de a IA ignorá-lo ou classificá-lo incorretamente.

A alta variabilidade estatística, com F_1 -scores oscilando entre 98% (espécie) e 36% (injeções), demonstra que a confiabilidade da IA é dependente da natureza do dado. Variáveis binárias ou com terminologia padronizada são altamente confiáveis; variáveis que exigem extração de tabelas ou inferência, como a de intervalos temporais, apresentam riscos. Assim, os resultados indicam que os modelos de IA ainda carecem da compreensão contextual e da numeracia necessária para substituir integralmente revisores humanos em domínios técnicos. Todavia, se modelos de IA conseguem sinalizar quais espécies estão sendo submetidas a quais fármacos com 98% e 95% de confiabilidade (F_1 -score), respectivamente, a redução do uso de animais prevista pelos 3Rs se torna uma meta mais alcançável, evitando-se a repetição de estudos cujas respostas já estão perdidas no “dilúvio de dados” da literatura científica.

6.6 Limitações e perspectivas futuras

Este estudo piloto apresenta limitações que devem ser consideradas para a interpretação dos achados e para o escalonamento da metodologia. A primeira limitação reside na extração automatizada ainda necessitar de um esforço de revisão humana que pode se aproximar do

tempo da extração manual, mitigando parte do ganho de eficiência esperado.

Outra limitação relevante se refere ao uso de interfaces de *chat* em detrimento de requisições via API. Embora a escolha por *chatbots* (ChatGPT e Gemini) tenha sido deliberada para garantir a acessibilidade a pesquisadores sem domínio de programação, essa abordagem introduz gargalos logísticos. Além disso, a natureza das atualizações desses modelos — o fenômeno conhecido como *model drift* — significa que o desempenho observado atualmente pode variar em versões futuras, exigindo que o protocolo de engenharia de *prompt* seja constantemente revalidado para manter os níveis de acurácia reportados.

No que tange aos dados quantitativos, a dependência de *scripts* em Python/R permanece como um desafio técnico. Os LLMs, isoladamente, ainda não são uma solução completa para revisões sistemáticas farmacológicas, mas sim um componente de um ecossistema computacional mais amplo que exige habilidades multidisciplinares do pesquisador.

Para o desdobramento desta pesquisa, vislumbra-se o desenvolvimento de uma interface de usuário especializada que integre os LLMs via API a um ambiente de verificação “humano-no-ciclo” (*human-in-the-loop*). Tal ferramenta permitiria que o pesquisador visualizasse, em tempo real, o trecho do PDF de onde a IA extraiu a informação, facilitando a auditoria e reduzindo drasticamente o tempo de conferência. Adicionalmente, o uso de modelos menores e especializados, como o PubMedBERT, com dicionários farmacológicos específicos e regras de extração tabular poderia resolver as falhas estruturais observadas.

Finalmente, a aplicação deste protocolo em bibliotecas massivas ($n > 1000$ artigos) é o próximo passo para medir o impacto real na redução e refinamento do uso de animais. Ao possibilitar que grandes grupos de pesquisa sintetizem décadas de evidências em poucas semanas, a automação baseada em IA tem o potencial de transformar a ciência pré-clínica, auxiliando que animais não sejam utilizados em experimentos cujas respostas já foram, estatisticamente, fornecidas pela literatura existente, mas que permaneciam invisíveis devido ao gargalo da síntese manual.

7 CONCLUSÃO

O presente estudo cumpriu seu objetivo geral ao desenvolver e validar um protocolo computacional baseado em inteligência artificial para a automação da extração de dados em revisões sistemáticas pré-clínicas. Os resultados obtidos permitem concluir que a integração de IA à ciência da síntese de evidências não é apenas uma possibilidade tecnológica, mas uma

necessidade metodológica para enfrentar o atual “dilúvio de dados” na área biomédica. Contudo, a padronização do protocolo baseado em IA proposta neste trabalho não se encerra na interação com a IA, mas estabelece que um protocolo de extração automatizada viável deve, obrigatoriamente, incluir uma etapa computacional de processamento de dados para garantir que cada unidade amostral seja corretamente identificada e validada, permitindo uma análise estatística justa e precisa.

A hipótese de que a aplicação de um protocolo baseado em LLMs permitiria a extração de dados com acurácia equivalente ao avaliador humano e eficiência temporal superior foi parcialmente confirmada. Em termos de eficiência temporal, a superioridade da IA foi absoluta, reduzindo o cronograma de extração de semanas para segundos. O uso do Gemini 2.5 Pro demonstrou que a janela de contexto massiva é um diferencial técnico que viabiliza a aplicação na prática acadêmica. No que diz respeito à acurácia e confiabilidade, a equivalência ao humano foi verificada parcialmente em variáveis descritivas e qualitativas (P do acrônimo PICO), como espécie e tipo de antidepressivo. Contudo, a hipótese foi refutada para variáveis quantitativas complexas e desfechos que exigem inferência lógica, como doses e regimes de intervenção, onde a IA apresentou taxas de omissão significativas.

A validação deste protocolo demonstra que a inteligência artificial pode atuar como um “revisor auxiliar”, capaz de realizar a mineração da literatura sob estrita supervisão humana. A abordagem híbrida preserva o rigor científico enquanto acelera drasticamente a produção de evidências. Do ponto de vista ético, este trabalho fundamenta o uso da tecnologia como uma ferramenta em prol do princípio dos 3Rs. Ao tornar as revisões sistemáticas mais rápidas e acessíveis, a automação permite que a comunidade científica identifique redundâncias experimentais de forma precoce, contribuindo diretamente para a redução do uso de animais em pesquisas cujas respostas já se encontram dispersas e escondidas no volume massivo da literatura científica.

7.1 Considerações finais

A transição da extração manual exaustiva para a curadoria automatizada sinaliza uma nova era na farmacologia baseada em evidências. Este estudo piloto buscou se alinhar aos estudos bases da área biomédica integrada à ciência de dados, com o objetivo compartilhado de que futuras ferramentas rompam o gargalo das revisões sistemáticas, garantindo que o conhecimento científico alcance a aplicação clínica com a velocidade que a saúde humana e o bem-estar animal exigem.

REFERÊNCIAS

ALSHAMI A., ELSAYED M., ALI E., ELTOUKHY A.E.E., ZAYED T. Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. **Systems**, v. 11, n. 351, 2023.

ANTHROPIC. Prompt engineering overview. **Claude Docs**, 2025. Disponível em: <https://platform.claude.com/docs/en/build-with-claude/prompt-engineering/overview>. Acesso em 10 de agosto de 2024.

BAHOR Z., LIAO J., CURRIE, G., AYDER C., MACLEOD M., MCCANN S. K., BANNACH-BROWN A., WEVER K., SOLIMAN N. WANG Q. DORAN-CONSTANT L., YOUNG L., SENA E. S., SENA C. Development and uptake of an online systematic review platform: the early years of the CAMARADES Systematic Review Facility (SyRF). **BMJ Open Science**, v. 5, n. 1, março de 2021.

BANNACH-BROWN A., PRZYBYŁA P., THOMAS J., RICE A. S. C., ANANIADOU S., LIAO J., MACLEOD M. R. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. **Systematic Reviews**, v. 8, n. 1, 2019.

BASTIAN H., GLASZIOU P., CHALMERS I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? **PLoS Medicine**, v. 7, n. 9, 2010.

BENDER E. M. et al. On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. **ACM**, pp. 610–623, 2021.

BEERY A. K., ZUCKER I. Sex bias in neuroscience and biomedical research. **Neuroscience Biobehavioral Reviews**, v. 35 n. 3, pp. 565-72, janeiro de 2011.

BOLZAN, J. A. **Teoria neurogênica dos efeitos dos antidepressivos sobre o comportamento: revisão sistemática e meta-análise**. 149 páginas. Dissertação (mestrado), Universidade Federal de Santa Catarina, Florianópolis, 2020.

BOLZAN J. A., OLIVEIRA C. L. A compact guide to the systematic review and meta-analysis of the literature in neuroscience. **Journal for Reproducibility in Neuroscience**, 2022.

BOUTRON I., CHAIMANI A., MEERPOHL J. J., HRÓBJARTSSON A., DEVANE D., RADA G., TOVEY D., GRASSELLI G., RAVAUD P. The COVID-NMA Project: Building an Evidence Ecosystem for the COVID-19 Pandemic. COVID-NMA Consortium. **Annals of Internal Medicine**, v. 173, n. 12, pp. 1015-1017, dezembro de 2020.

BORJI A. A Categorical Archive of ChatGPT Failures. **ArXiv preprint**, 2023. <arXiv:2302.03494v8>.

BROWN S. Machine learning explained. **MIT Sloan**, 2021. Disponível em: <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>. Acesso em 10 de dezembro de 2025.

CAO C., ARORA R., CENTO P., MANTA K., FARAHANI E., CECERE M., SELEMON A., SANG J., et al. Automation of Systematic Reviews with Large Language Models. **MedRxiv**, 2025. <2025.06.13.25329541>.

CHAMULEAU S. A. J., VAN DER NAALD M., CLIMENT A. M., KRAAIJEVELD A. O., WEVER K. E., DUNCKER D. J., FERNÁNDEZ-AVILÉS F., BOLLI R. Transnational Alliance for Regenerative Therapies in Cardiovascular Syndromes (TACTICS) Group. Translational Research in Cardiovascular Repair: A Call for a Paradigm Shift. **Circulation Research**, v. 122, n. 2, pp. 310-318, 2018.

CHRISTEN P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. **Springer Science & Business Media**, 2012.

CLARK J., GLASZIOU P., DEL MAR C., BANNACH-BROWN A., STEHLIK P., SCOTT A. M. A full systematic review was completed in 2 weeks using automation tools: a case study. **Journal of Clinical Epidemiology**, v. 121, pp. 81-90, 2020.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, v. 20, n. 1, pp. 37-46, 1960.

DAVIDSON M., KORFITSSEN C. B., RIVEROS C., CHAIMANI A., BOUTRON I. POST-publication peer review and the identification of methodological and reporting issues in COVID-19 trials: a qualitative study. **BMJ Evidence Based Medicine**, v. 30, n. 4, pp. 233-240, julho de 2025.

DEEKS J. J., HIGGINS J. P. T., ALTMAN D. G. Chapter 10: Analysing data and undertaking meta-analyses. In: HIGGINS J.P.T., THOMAS J., CHANDLER J., CUMPSTON M., LI T., PAGE M. J., WELCH V. A. (editores). **Cochrane Handbook for Systematic Reviews of Interventions**, v. 6.4 (atualizado em agosto de 2024). Cochrane, 2024. Disponível em: www.training.cochrane.org/handbook. Acesso em 01 de setembro de 2024.

DIOGENES M. From the Bridge: Chatbot Confidential: Machine Magic and the Power of Language. **Stanford Teaching Writing**, 2023. Disponível em: <https://teachingwriting.stanford.edu/news/bridge-chatbot-confidential-machine-magic-and-power-language>. Acesso em 10 de dezembro de 2025.

ELBADAWI M., LI H., BASIT A. W., GAISFORD S. **The role of artificial intelligence in generating original scientific research**. International Journal of Pharmaceutics, v. 652, p. 123741, março de 2024.

EMAMI M., SHIRANI M. Comparing the performance of ChatGPT, DeepSeek, and Gemini in systematic and umbrella review tasks over time. **Journal of the American Dental Association**, v. 156, n. 12, pp. 1014-1023, 2025.

GARRITTY C., GARTLEHNER G., NUSSBAUMER-STREIT B., KING V. J., HAMEL C., KAMEL C., AFFENGRUBER L., STEVENS A. Cochrane Rapid Reviews Methods Group offers evidence-informed guidance to conduct rapid reviews. **Journal of Clinical Epidemiology**, v. 130, pp. 13-22, fevereiro de 2021.

GARTLEHNER G., KAHWATI L., HILSCHER R., THOMAS I., KUGLEY S., CROTTY K., et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. **Research Synthesis Methods**, v. 15 n. 4, pp. 576-589, 2024.

GEVA M., GUPTA A., BERANT J. Injecting Numerical Reasoning Skills into Language Models. **ArXiv preprint**, 2020. <arXiv:2004.04487v1>

GRANDINI M., BAGLI E., VISANI G. Metrics for multi-class classification: an overview. **ArXiv preprint**, 2020. <arXiv:2008.05756>.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J., POON H. Domain-specific language model pretraining: case study on biomedical NLP. **ACM Transactions on Computing for Healthcare**, v. 3, n. 1, pp. 2:1-23, 2021.

GURURAJAN A., REIF A., CRYAN J. F., SLATTERY D. A. **The future of rodent models in depression research.** *Nature Reviews Neuroscience*, v. 20, pp. 686-701, novembro de 2019.

GUTIERREZ B. J., MCNEAL N., WASHINGTON C., CHEN Y., LI L., SUN H., SU Y. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. In: Findings of the Association for Computational Linguistics: NAACL. **Association for Computational Linguistics**, pp. 4526–4541, 2022.

HAMZE F. C. **Avaliação de risco de viés em estudos relacionando a neurogênese hipocampal à ação dos antidepressivos.** 53 páginas. Trabalho de Conclusão de Curso (graduação), Universidade Federal de Santa Catarina, Florianópolis, 2023.

HIGGINS J. P. T., THOMAS J., CHANDLER J., CUMPSTON M., LI T., PAGE M. J., et al. **Cochrane Handbook for Systematic Reviews of Interventions**, versão 6.5 (atualizada em agosto de 2024). Cochrane, 2024. Disponível em www.cochrane.org/handbook. Acesso em fevereiro de 2025.

HOFFMANN F., ALLERS K., ROMBEY T., HELBACH J., HOFFMANN A., MATHES T., PIEPER D. 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**, v. 138, pp. 1-11, junho de 2021.

HOFFMANN T. C., OXMAN A. D., IOANNIDIS J. P. et al. Enhancing the usability of systematic reviews: a survey of methodologists and end users. **BMJ Evidence-Based Medicine**, v. 26, n. 3, p. 108, 2021.

HOLTZMAN A., WEST P., SHWARTZ V., CHOI Y., ZETTLEMOYER L. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right **ArXiv preprint**, 2021. < arXiv:2104.08315v9>

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science & Engineering**, v. 9, n. 3, p. 90-95, 2007.

IBRAHIM E. C., GORGIEVSKI V., ORTIZ-TEBA P., BELZEAUX R., TURECKI G., et al. **Transcriptomic Studies of Antidepressant Action in Rodent Models of Depression: A First Meta-Analysis.** *International Journal of Molecular Science*, v. 23, n.21,

p. 13543, novembro de 2022.

INEICHEN B. V., ROSSO M., MACLEOD M. R. From data deluge to publomics: How AI can transform animal research. **Lab Animal**, v. 52, pp. 213–214, 2023.

INEICHEN B.V., HELD U., SALANTI G. et al. Systematic review and meta-analysis of preclinical studies. **Nature Review Methods**, v. 4, n. 72, 2024.

IOANNIDIS J. P. A. Why Most Clinical Research Is Not Useful. **PLoS Medicine** v. 13, n. 6, 2016.

JONNALAGADDA, S. R., GOYAL, P., & HUFFMAN, M. D. Automating data extraction in systematic reviews: a systematic review. **Systematic Reviews**, v. 4, n. 78, 2015.

KADAVATH S. CONERLY T., ASKELL A., HENIGHAN T., DRAIN D., et al. Language Models (Mostly) Know What They Know. **ArXiv preprint**, 2022. <arXiv:2207.05221v4>.

KANDPAL K., DENG H., ROBERTS A., WALLACE E., RAFFEL C. Large Language Models Struggle to Learn Long-Tail Knowledge. **ArXiv preprint**, 2022. <arXiv:2211.08411v2>.

KAPLAN B. L. F., HOBERMAN A. M., SLIKKER W. JR, SMITH M. A., et al. **Protecting Human and Animal Health: The Road from Animal Models to New Approach Methods**. Pharmacological Reviews, v. 76, n. 2, pp. 251-266, fevereiro de 2024.

KAPLAN J., MCCANDLISH S., HENIGHAN T., BROWN T. B., CHESS B., CHILD R., et al. **Scaling laws for neural language models**. **ArXiv Preprint**, 2020. <arXiv:2001.08361>.

KILKENNY C., BROWNE W. J., CUTHILL I. C., EMERSON M., ALTMAN D. G. Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. **PLoS Biology** v. 8, n. 6, junho de 2010.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, pp. 159–174, 1977.

LANDHUIS, E. Scientific literature: Information overload. **Nature** v. 535, pp. 457–

458, 2016.

LI J., SUN A., HAN J., LI C. A Survey on Deep Learning for Named Entity Recognition. **ArXiv Preprint**, 2018. < arXiv:1812.09449v3>.

LIU, N. F., LIN K., HEWITT J., PARANJAPPE A., et al. Lost in the Middle: How Language Models Use Long Contexts. **Transactions of the Association for Computational Linguistics**, v. 12, p. 157-173, 2024.

MACLEOD M. R., LAWSON MCLEAN A., KYRIAKOPOULOU A., SERGHIOU S., DE WILDE A., SHERRATT N., et al. Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. **PLoS Biology** v. 13, n. 10, outubro de 2015.

MACLEOD M. R., MICHIE S., ROBERTS I., DIRNAGIL U., et al. Biomedical research: increasing value, reducing waste. **The Lancet**, v. 383, n. 9912, p. 101-104, 2014.

MARSHALL I. J., KUIPER J., WALLACE B. C. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. **Journal of the American Medical Informatics Association**, v. 23, n. 1, pp. 193-201, janeiro de 2016.

MARSHALL I. J., NYE B., KUIPER J., NOEL-STORR A., MARSHALL R., MACLEAN R., et al. Trialstreamer: A living, automatically updated database of clinical trial reports. **Journal of the American Medical Informatics Association**, v. 27, n. 12, pp. 1903-1912, dezembro de 2020.

MARSHALL I. J., WALLACE B.C. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. **Systematic Reviews**, v. 8, n. 163, 2019.

MCKINNEY W. Data structures for statistical computing in python. **Proceedings of the 9th Python in Science Conference**, Texas, v. 445, pp. 51–56, 2010.

MENON J. M. L., RITSKES-HOITINGA M., POUND P., VAN OORT E. The impact of conducting preclinical systematic reviews on researchers and their research: A mixed method case study. **PLoS ONE**, v. 16, n. 12, dezembro de 2021.

MESSERI L., CROCKETT M. J. **Artificial intelligence and illusions of understanding in scientific research**. *Nature*, v. 627, pp. 49–58, março de 2024.

MOHER D., HOPEWELL S., SCHULZ K. F., MONTORI V., et al. CONSORT 2010 Explanation and Elaboration: Updated Guidelines for Reporting Parallel Group Randomised Trials. **BMJ**, v. 340, c1232, 2010.

MUKHERJEE P., ROY, S., GHOSH D., NANDI S. K. **Role of animal models in biomedical research: a review**. Laboratory Animal Research, v. 38, n. 18, julho de 2022.

MURAD M. H., ASI N., ALSAWAS M., ALAHDAB F. New evidence pyramid. **Evidence Based Medicine**, v, 21, n. 4, pp. 125-7, 2016.

NASA. What is Artificial Intelligence? **National Aeronautics and Space Administration (NASA)**, 2024. Disponível em: <https://www.nasa.gov/what-is-artificial-intelligence/>. Acesso em 10 de dezembro de 2025.

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI). PubMed. **National Library of Medicine**, 1988. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/>. Acesso em 10 de dezembro de 2025.

NATURE PORTFOLIO. **Nature Collection**. Nobel Prize in Chemistry, 2024. Disponível em: <https://www.nature.com/collections/edjcfdihi>. Acesso em 10 de dezembro de 2025.

NAVEED H. et al. A comprehensive overview of large language models. **ArXiv preprint**, 2023. <arXiv:2307.06435>.

NORI H., KING N., MCKINNEY S. M., CARIGNAN D., HORVITZ E. Capabilities of GPT-4 on Medical Challenge Problems. **ArXiv preprint**, 2023. <arXiv:2303.13375>.

O'MARA-EVES A., THOMAS J., MCNAUGHT J., MIWA M., ANANIADOU S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. **Systematic Reviews**, v. 4, n. 5, 2015.

OUZZANI M., HAMMADY H., FEDOROWICZ Z., ELMAGARMID A. Rayyan-a web and mobile app for systematic reviews. **Systematic Reviews**, v. 5, n. 1, p. 210, 2016.

PEDREGOSA F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, pp. 2825–2830, 2011.

PERAKAKIS N., YAZDANI A., KARNIADAKIS G. E., MANTZOROS C. Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. **Journal of Metabolism**, v. 87, A1-A9, 2018.

PERCIE DU SERT N., HURST V., AHLUWALIA A., ALAM S., AVEY M. T., BAKER M., et al. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. **PLOS Biology**, v. 18, n. 7, 2020.

PEREL P., ROBERTS I., SENA E., WHEBLE P., BRISCOE C., SANDERCOCK P., MACLEOD M., et al. Comparison of treatment effects between animal experiments and clinical trials: systematic review. **BMJ**, v. 334, n. 7586, p. 197, 2007.

POSIT TEAM. RStudio: Integrated Development Environment for R. **Posit Software**, PBC, Boston, MA, 2025. Disponível em: <http://www.posit.co/>. Acesso em 10 de dezembro de 2025.

POWERS, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. **Journal of Machine Learning Technologies**, v. 2, n. 1, pp. 37-63, 2011.

PRESS RELEASE. **Nobel Prize Outreach**, 2025. Disponível em: <https://www.nobelprize.org/prizes/chemistry/2024/press-release/>. Acesso em 10 de dezembro de 2025.

PROSPERO. PROSPERO: International prospective register of systematic reviews. **Centre for Reviews and Dissemination**, University of York, 2025. Disponível em: <https://www.crd.york.ac.uk/prospero/>. Acesso em 15 de dezembro de 2025.

PYTHON SOFTWARE FOUNDATION. **Python Language Reference**, version 3.x. Disponível em: <https://www.python.org>. Acesso em 10 de dezembro de 2025.

QURESHI R., SHAUGHNESSY D., KA R. et al. AI in systematic reviews: The promises and perils of ChatGPT. **Journal of Clinical Epidemiology**, v. 164, pp. 128-132, 2023

R CORE TEAM. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria, 2025. Disponível em: <https://www.R-project.org/>. Acesso em 10 de dezembro de 2025.

RITSKES-HOITINGA M, VAN LUIJK J. How Can Systematic Reviews Teach Us More about the Implementation of the 3Rs and Animal Welfare? **Animals**, v. 9, n. 12, p. 1163, 2019.

RITSKES-HOITINGA M., LEENAARS M., AVEY M., ROVERS M., SCHOLTEN R. Systematic reviews of preclinical animal studies can make significant contributions to health care and more transparent translational medicine. **Cochrane Database of Systematic Reviews**, v. 2014, n. 3, 2014.

RUSSEL W. M. S., BURCH R. L. **The Principle of Humane Experimental Technique**. 1. ed. Londres: Methuen & CO, 1959.

RUSSELL S. J., NORVIG P. **Artificial intelligence: a modern approach**. Pearson, 4ª edição, 2020.

SAPAROV A., HE H. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. **ArXiv preprint**, 2022. < arXiv:2210.01240v4>

SCIKIT-LEARN DEVELOPERS. Model selection and evaluation. **User Guide**, seção 3, 2025. Disponível em: https://scikit-learn.org/stable/model_selection.html. Acesso em 8 junho de 2025.

SEARLE J. Minds, brains, and programs. **Behavioral and Brain Sciences**, v. 3, n. 3, pp. 417-424, 1980.

SEYHAN A. A. Lost in translation: the valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. **Translational Medicine Communications** v. 4, n. 18, 2019.

SHOJANIA K. G., SAMPSON M., ANSARI M. T., JI J., DOUCETTE S., MOHER D. How quickly do systematic reviews go out of date? A survival analysis. **Annals of Internal Medicine**, v. 147, n. 4, pp. 224-33, 2007.

SOBOCZENSKI F., TRIKALINOS T. A., WHITING P. Machine learning for data extraction in systematic reviews: the challenge of generalizability. **Research Synthesis Methods**, v. 10, n. 3, pp. 405-415, 2019.

SYSPHAR. Systematic Pharmacology. **CAMARADES Brazil**, 2022. Disponível em

<https://camaradesbrasil.bio.br/sysphar>. Acesso em 10 de dezembro de 2025.

TANG R., HAN X., JIANG X., HU X. Does Synthetic Data Generation of LLMs Help Clinical Text Mining? **ArXiv preprint**, 2023. <arXiv:2303.04360v2>.

TIDYVERSE DEVELOPERS. **Tidyverse**. Disponível em: <https://tidyverse.org/>. Acesso em 10 de março de 2025.

THAWANI A., PUJARA J., SZEKELY P. A., ILIEVSKI F. Representing Numbers in NLP: a Survey and a Vision. **ArXiv preprint**, 2021. <arXiv:2103.13136v1>

THE PANDAS DEVELOPMENT TEAM. **pandas-dev/pandas: Pandas 2.2.x**. 2024. Disponível em: <https://pandas.pydata.org>. Acesso em 10 de março de 2025.

TSOUMAKAS G., KATAKIS I. Multi-label classification: An overview. **International Journal of Data Warehousing and Mining**, v. 3, n. 3, pp. 1-13, 2007.

USHEY K., ALLAIRE J. J., TANG Y. Reticulate: Interface to 'Python'. R package version 1.44.1. **Github**, 2025. Disponível em: <https://rstudio.github.io/reticulate/>. Acesso em 10 de março de 2025.

VASWANI A., SHAZEER N., PARMAR N. et al. Attention is all you need. **Advances in Neural Information Processing Systems**, v. 30, 2017.

VERITAS HEALTH INNOVATION. Covidence systematic review software. **Veritas Health Innovation**, 2014. Disponível em: <https://www.covidence.org/>. Acesso em 15 de dezembro de 2025.

VIRTANEN P., GOMMERS R., OLIPHANT T. E., HABERLAND M., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. **Nature Methods**, v. 17, n. 3, pp. 261-272, 2020.

VOELKL B., ALTMAN N. S., FORSMAN A., FORSTMEIER W., GUREVITCH J., et al. Reproducibility of animal research in light of biological variation. **Nature Review Neuroscience**, v. 21, n. 7, pp. 384-393, 2020.

VON MÜCKE-HEIM I. A., URBINA-TREVIÑO L., BORDES J., RIES C., SCHMIDT M. V., DEUSSING J. M. **Introducing a depression-like syndrome for translational**

neuropsychiatry: a plea for taxonomical validity and improved comparability between humans and mice. *Molecular Psychiatry*, v. 28, pp. 329–340, setembro de 2023.

WALLACE E., WANG Y., LI S., SINGH S., GARDNER M. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. ArXiv preprint, 2019. < arXiv:1909.07940v2>

WANG P., NAIR M. S., LIU L., IKETANI S., LUO Y., GUO Y., WANG M., et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. **Nature**, v. 593, n. 7857, pp. 130-135, maio de 2021.

WANG Q., TIMBERLAKE M. A., PRALL K., DWIVEDI Y. The recent progress in animal models of depression. **Progress in Neuro-Psychopharmacology & Biological Psychiatry**, v. 77, pp. 99-109, julho de 2017.

WANG S., SCELLS H., KOOPMAN B., ZUCCON, G. Can ChatGPT Draft Systematic Review Protocols? A Case Study. **SIGIR**, 2024.

WASKOM M. L. **Seaborn Tutorial**, 2024. Disponível em: <https://seaborn.pydata.org/tutorial.html>. Acesso em 5 de abril de 2025.

WASKOM M. L. Seaborn: statistical data visualization. **Journal of Open-Source Software**, v. 6, n. 60, p. 3021, 2021.

WEI J., TAY Y., BOMMASANI R., RAFFEL C., ZOPH B., BORGEAUD S., et al. Emergent abilities of large language models. **Transactions on Machine Learning Research**, 2022.

WICKHAM H. Welcome to the Tidyverse. **Journal of Open-Source Software**, v. 4, n. 43, p. 1686, 2019.

WICKHAM H. Ggplot2: Elegant Graphics for Data Analysis. **Springer-Verlag**, 2016.

WICKHAM H. Tidy Data. **Journal of Statistical Software**, v. 59, n. 10, pp. 1–23, 2014.

WOITOWICH N. C., BEERY A., WOODRUFF T. A 10-year follow-up study of sex inclusion in the biological sciences. **Elife**, v. 9, p. e56344, junho de 2020.

WOLTERS KLUWER. Ovid. **Wolters Kluwer**, 1992. Disponível em: <https://ovidsp.ovid.com/>. Acesso em 15 de dezembro de 2025.

ZHAO, W. X., ZHOU K., LI J., TANG T., WANG X., et al. A survey of large language models. **ArXiv preprint**, 2023. <arXiv:2303.18223>.

ZHAO T. Z., WALLACE E., FENG S., KLEIN D., SINGH S. Calibrate Before Use: Improving Few-Shot Performance of Language Models. **ArXiv preprint**, 2021. <arXiv:2102.09690v2>.