Viés em Sistemas de Informação:

Uma Revisão Sistemática sobre Identificação, Mitigação e Governança Ética.

Autores: Diogo Lima, Gabriel Aragão, Guilherme Lopes, Mader Gabriel, Matheus Braga, Maria Clara Barretto, Luiz Eduardo Schmalz, Luiz Roberto.

Contextualização e Problema

Contexto:
Sistemas de Informação (SI) baseados em IA apoiam decisões críticas (saúde, justiça criminal, contratações).

O Problema:
O **Viés Algorítmico** gera erros sistemáticos que ampliam desigualdades sociais.

Exemplos reais: Discriminação em reincidência criminal e filtros de contratação com viés de gênero.

Abordagens puramente técnicas são insuficientes para um problema sociotécnico.

Objetivo Geral:

Analisar metodologias de identificação e mitigação de viés, propondo um framework ético-técnico baseado em justiça (fairness), transparência e responsabilização (accountability).

Objetivos Específicos:

1. Mapear taxonomias de viés no ciclo de vida dos dados.
2. Avaliar métodos e métricas de detecção.
3. Analisar estratégias de mitigação e seus trade-offs.
4. Operacionalizar princípios éticos em diretrizes de engenharia.

Protocolo de Busca

Bases de Dados:
ACM Digital Library, IEEE Xplore, Scopus e SpringerLink.

Período:
2020 a 2025 (Foco em avanços recentes em IA Responsável)

String de Busca (Exemplo):

("algorithmic bias" OR "bias mitigation") AND ("information systems" OR "machine learning") AND ("ethics" OR "fairness")

Protocolo de Busca

Critérios de Inclusão (IC):

- Artigos completos
- Revisados por pares
- Em inglês
- Área de Computação/SI

Critérios de Exclusão (EC):

- Papers curtos
- Anteriores a 2020
- Teses, revisões secundárias
- Duplicatas entre bases de busca.
- Artigos sem acesso completo ou pagos.

Seleção e Qualidade

Inicialmente: 367 artigos.

Selecionados ao final: 49 estudos primários de alta qualidade.

Critérios de Qualidade (QC):
Avaliação feita com base em 5 critérios centrais (Objetivos claros, Metodologia apropriada, Métodos de viés, Implicações éticas, Conclusões suportadas)

Uso de LLMs com "Human-in-the-loop”:
Foram utilizados ChatGPT-4 e Gemini 1.5 Pro nas etapas de triagem e extração de informações.

Validação humana obrigatória para resolver divergências (aprox. 30% de divergência requerendo intervenção humana).

Perguntas de Pesquisa

Pergunta 1: Quais são as principais taxonomias de viés documentadas na literatura de Sistemas de Informação e em quais etapas do ciclo de vida elas são introduzidas?

Pergunta 2: Quais métodos, técnicas e métricas são utilizados para avaliar viés e equidade em modelos de Sistemas de Informação?

Pergunta 3: Como as estratégias de mitigação impactam o trade-off entre desempenho do sistema, equidade e transparência?

Pergunta 4: Como princípios éticos podem ser operacionalizados em diretrizes práticas para a gestão de vieses?

RQ1 - Taxonomias e Ciclo de Vida

Onde o viés entra? O viés não é apenas um problema de dados, mas de todo o ciclo.

Coleta/Preparação:
Viés de Representação, Estrutural e de Medição (dados não refletem a população ou refletem desigualdades históricas).

Treinamento:
Viés Algorítmico (amplificado pela arquitetura do modelo).

Operação/Uso:
Viés Dinâmico (quando o contexto de uso difere do treinamento/data drift).

RQ2 - Métodos e Métricas de Avaliação

Como medir? Acurácia global esconde discriminação contra subgrupos.

Avaliação Estratificada por Grupo: Medir performance separada para grupos sensíveis.

Análise de Erro Diferencial: Comparar taxas de falsos positivos/negativos entre grupos.

Testes Contrafatuais: Para LLMs, testar como a saída muda ao alterar apenas o atributo protegido

Métricas Chave: Paridade Demográfica, Igualdade de Oportunidades.

RQ3 - Mitigação e Trade-offs

Como corrigir? Intervenções em três estágios:

Pré-processamento: Rebalanceamento dos dados (Corrige a desproporção entre grupos, mas pode eliminar informações úteis).

In-processing: O algoritmo é treinado para evitar discriminação, fairness(Mais robusto, maior custo computacional).

Pós-processamento: Recalibragem da decisão (Altera a régua de aprovação, risco de mascarar o problema real).

Fairness vs. Performance: Ganho em justiça muitas vezes custa uma pequena redução na acurácia global.

RQ4 - Operacionalização da Ética

Como governar? Transformar princípios éticos em práticas técnicas e organizacionais.

Fairness (Justiça):
Validação por subgrupos (Testes estratificados e Design Inclusivo).

Transparency (Transparência): Rastreabilidade via Logs detalhados e documentação (Model Cards).

Accountability (Responsabilidade)
Governança institucional (Auditoria contínua e Comitês de Ética).

Resultados e discussões

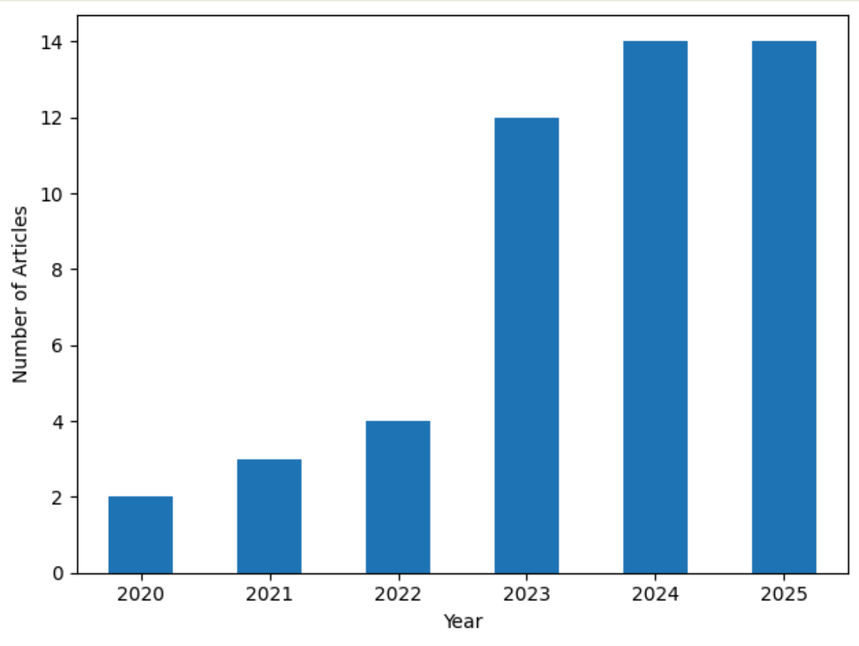
Avaliação da Qualidade dos Estudos:
49 artigos avaliados com a metodologia de pontuação, com média de 3.82 e mediana de 4.0

Cerca de 55% dos estudos atingiram pontuação 4.0 ou mais, indicando alta qualidade metodológica, especialmente em objetivos claros e contribuições sobre viés.

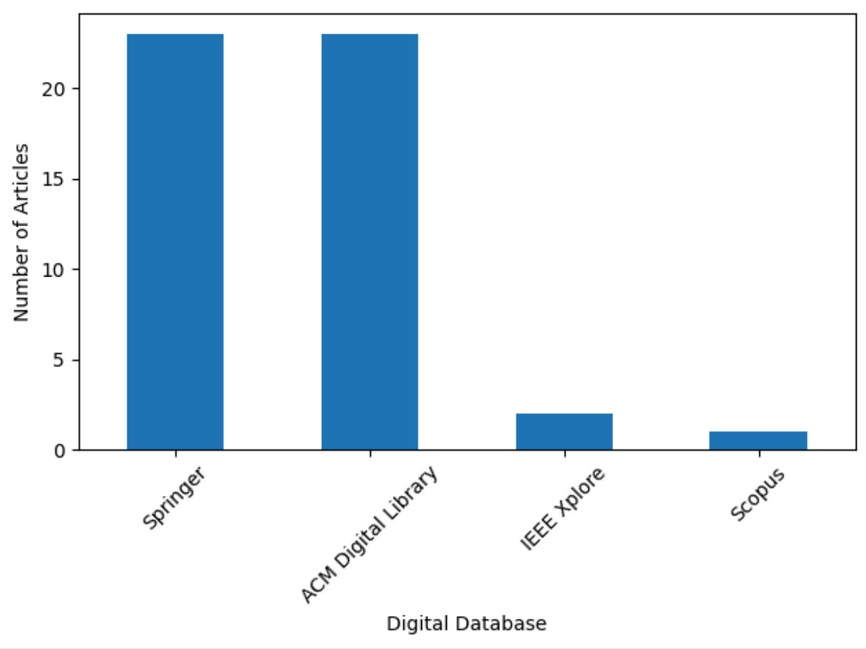
Modelo LLM (ChatGPT vs Gemini):
69.4% de consenso em extrações de dados entre os dois modelos.
30.6% divergente, necessitando validação humana para casos mais complexos.

Resultados e discussões

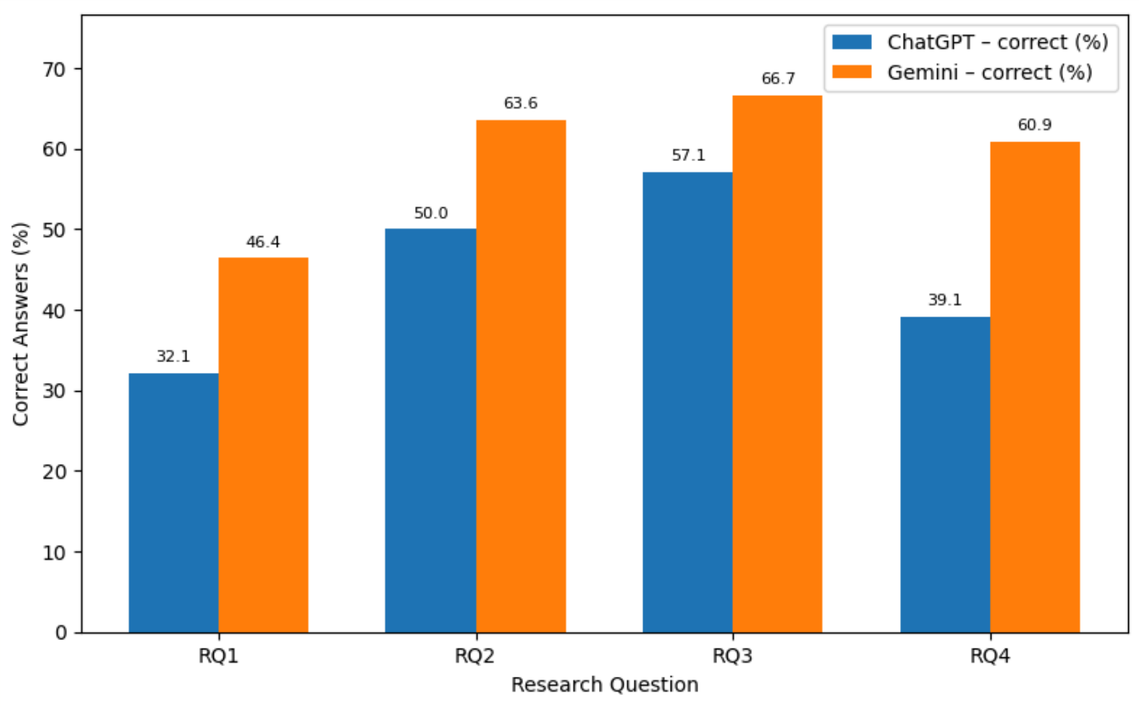
Distribuição dos Artigos:
Publicações concentradas nos anos 2023–2025, com maioria em periódicos e canais de conferências.



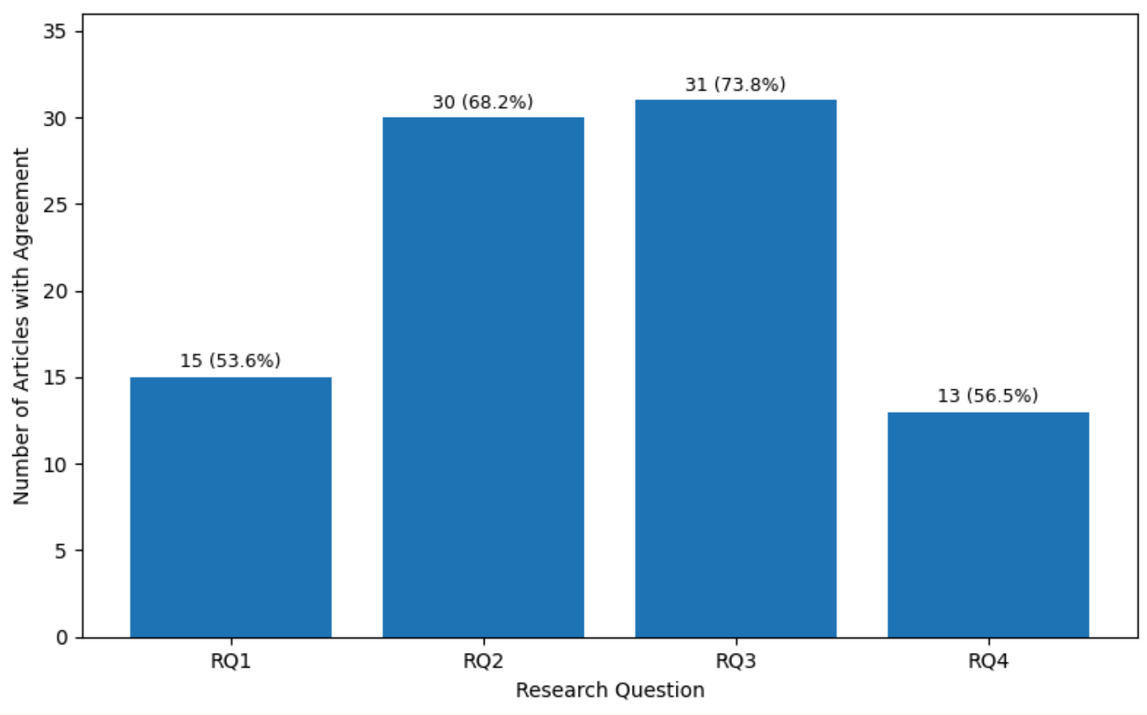
Principais bancos de dados:
ACM Digital Library e SpringerLink.



Resultados e discussões

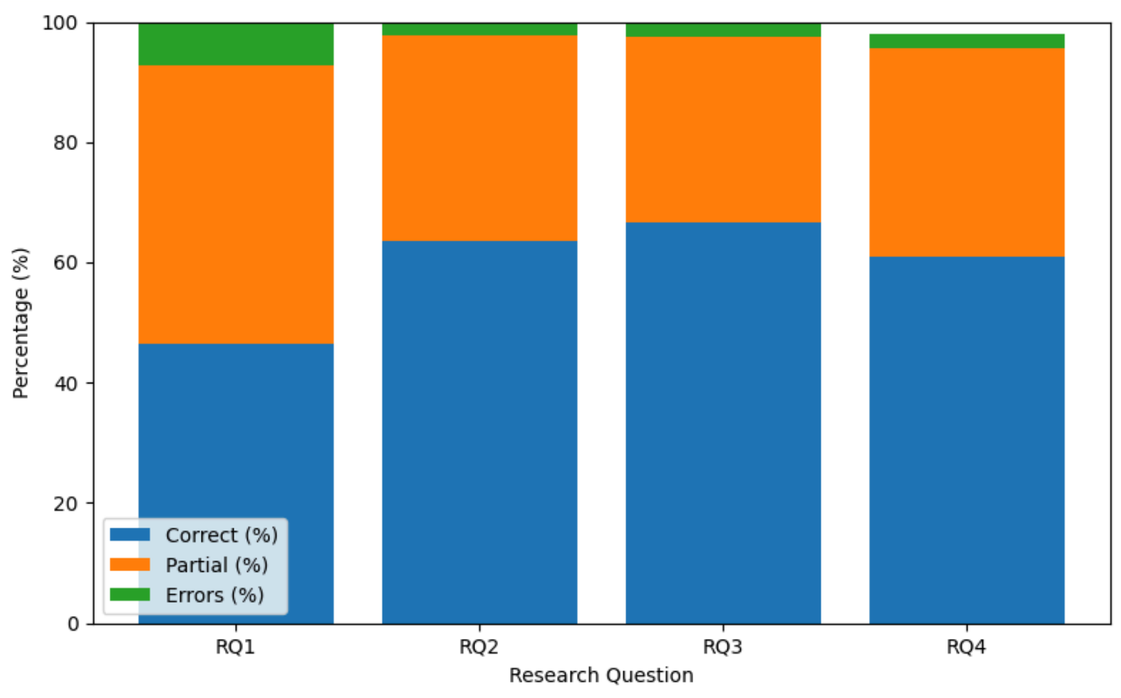


Accuracy comparison (rate of correct answers) between ChatGPT and Gemini for each of the four research questions (RQs)

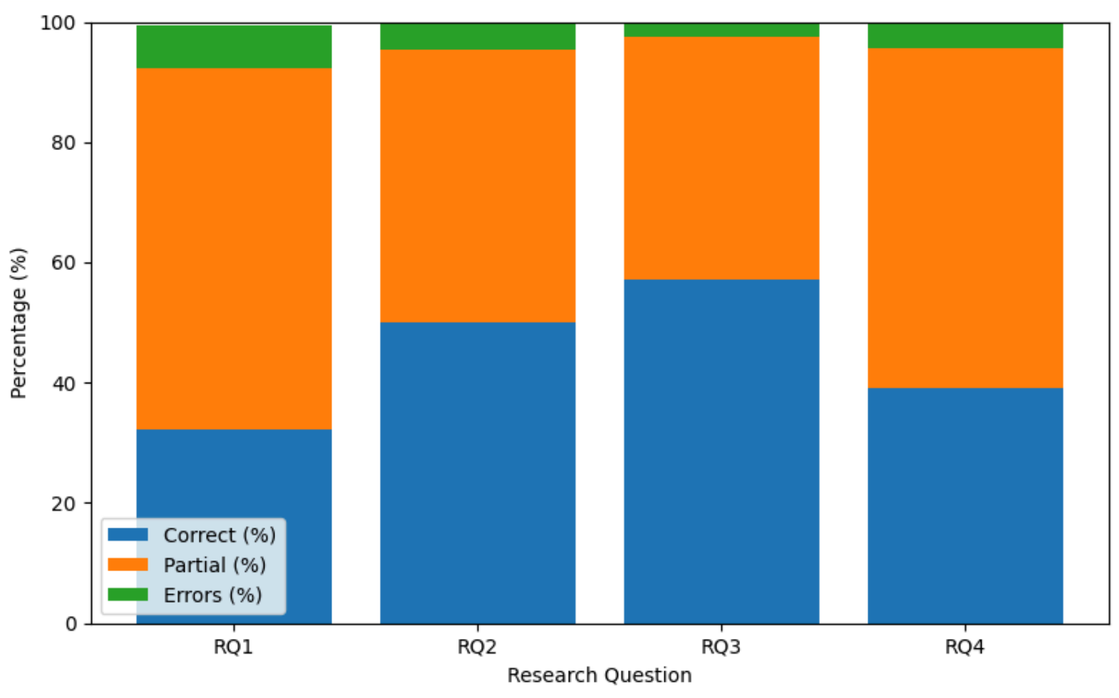


Articles per Research Question (RQ)

Resultados e discussões



Gemini – Outcomes per RQ



ChatGPT – Outcomes per RQ

Conclusões e Trabalhos Futuros

Por fim, este trabalho evidenciou que o viés algorítmico constitui um desafio sociotécnico complexo, presente em todas as fases do ciclo de vida dos dados, desde a coleta até a implantação dos sistemas, impactando diretamente a tomada de decisão automatizada.

A revisão sistemática proporcionou uma visão abrangente das causas, formas de avaliação e estratégias de mitigação do viés, destacando as limitações de métricas tradicionais e a necessidade de análises estratificadas e auditorias contínuas.

Conclusões e Trabalhos Futuros

Os resultados reforçam que princípios éticos, como justiça, transparência e responsabilização, devem ser incorporados como requisitos técnicos, superando a abordagem de soluções isoladas e promovendo arquiteturas orientadas à auditoria e ao design inclusivo.

Como direções futuras, destaca-se a importância da integração entre soluções técnicas e estruturas de governança, assegurando que Sistemas de Informação que mediam decisões críticas sejam desenvolvidos de forma justa, transparente e responsável.

Referências

01 **Androutsopoulou, M., Gkotsis, G., Gkioulos, V., & Douligeris, C. (2025). Towards AI-Enabled Cyber-Physical Infrastructures—Challenges, Opportunities, and Implications for a Data-Driven eGovernment Theory, Policy, and Practice. Journal of the Knowledge Economy._**

02 Bano, M., Ali, S., & Zowghi, D. (2025a). Envisioning responsible quantum software engineering and quantum artificial intelligence. Automated Software Engineering._

03 Bano, M., Ali, S., & Zowghi, D. (2025b). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. Automated Software Engineering._

04 Esmaeilzadeh, P. (2025). Ethical implications of using general-purpose LLMs in clinical settings: a comparative analysis of prompt engineering strategies and their impact on patient safety. BMC Medical Informatics and Decision Making._

05 Foalem, P. L., Ben-Attou, R., Madi, E. A., & Bounouar, O. (2025). Logging requirement for continuous auditing of responsible machine learning-based applications. Empirical Software Engineering._

Referências

06 Gao, X., Shen, C., Jiang, W., Lin, C., Li, Q., Wang, Q., Li, Q., & Guan, X. (2024). Fairness in machine learning: definition, testing, debugging, and application. Science China Information Sciences, 67(10), 108101._

---

07 Kurumayya, V. (2025). Towards fair AI: a review of bias and fairness in machine intelligence. Journal of Computational Social Science._

---

08 Panarese, P., Grasso, M. M., & Solinas, C. (2025). Algorithmic bias, fairness, and inclusivity: a multilevel framework for justice-oriented AI. AI & SOCIETY._

- https://github.com/gabrielaragao01/Bias-in-Information-Systems-A-Systematic-Review-of-Identification-Mitigation-and-Ethical-Governance