Fichamento dos Relatórios de Resumo e Discussão dos Artigos Vistos.

Questão escolhida: **How can the collapse of generative AI models affect the accuracy and quality of responses provided by autonomous agents in interactions with human users in the corporate environment?**	
Frases escolhidas para consulta:	Collapse of generative AI models; Quality of AI responses; Autonomous AI agents; Autonomous AI agents in corporate applications; Impact of AI model degradation; Synthetic distributed data; Multi-agents in corporate applications; Bias in Multi-Agents; AI ethics and hallucinations; Agents AI in recommendation; Hallucinations in LLMs; Poisoning in data distribution for LLMs; Collapse models improve bias;
Fontes (revistas, conferências, etc) de artigos usados:	Nature; International Journal of Science and Research Archive; Frontiers of Computer Science; arXiv; IEEE; Electronic Commerce Research; Scientific Reports;
Nome dos artigos selecionados:	AI models collapse when trained on recursively generated data
	Review of autonomous systems and collaborative AI agent frameworks
	A Survey on Large Language Model based Autonomous Agents
	Can We Trust AI Agents? An Experimental Study Towards Trustworthy LLM-Based Multi-Agent Systems for AI Ethics

	Transforming Competition into Collaboration: The Revolutionary Role of Multi-Agent Systems and Language Models in Modern Organizations
	Transforming Competition into Collaboration: The Revolutionary Role of Multi-Agent Systems and Language Models in Modern Organizations
	Multi-Agent Large Language Models for Conversational Task-Solving
	Strong Model Collapse
	How to Synthesize Text Data without Model Collapse?
	The Impact of Large Language Models in Academia: from Writing to Speaking
	Mitigating Social Bias in Large Language Models: A Multi-Objective Approach Within a Multi-Agent Framework
	Towards Implicit Bias Detection and Mitigation in Multi-Agent LLM Interactions
	Fairness in Multi-Agent AI: A Unified Framework for Ethical and Equitable Autonomous Systems
	Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey
	Personalized Recommendation Systems using Multimodal, Autonomous, Multi Agent Systems

	Consumer reactions to technology in retail: choice uncertainty and reduced perceived control in decisions assisted by recommendation agents
	Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception
	Bias of AI-generated content: an examination of news produced by large language models
	A Multi-Agent Conversational Recommender System
	Poisoning and Backdooring Contrastive Learning

Resumo Geral

Esses estudos mostram como os modelos LLMs e os sistemas multiagentes estão influenciando cada vez mais a tecnologia, especialmente em áreas como recomendação, ética e desempenho. Um dos principais pontos discutidos é o "colapso do modelo", que acontece quando a IA é treinada com dados não curados e sintéticos, o que faz com que ela perca a diversidade dos dados e prejudique sua performance. Isso também pode aumentar os vieses nos modelos, o que é preocupante, especialmente em áreas como tomada de decisão e recomendação de produtos.

No caso dos sistemas multiagentes, muitos estudos exploram como esses agentes podem melhorar a interação com os usuários, ajudando a controlar o fluxo das conversas e coletar feedback para ajustar as respostas dos modelos. Esses sistemas têm mostrado potencial em vários setores, como comércio eletrônico, finanças e saúde, ao personalizar experiências e melhorar a adaptação a diferentes contextos.

Outro grande tema é o viés nos modelos de IA, principalmente de gênero e raça. Isso tem sido um desafio, e várias abordagens estão sendo testadas para reduzir esses vieses, como o uso de múltiplos agentes e ajustes nos prompts que são dados aos modelos. Também há discussões sobre como lidar com a ética em IA, criando

sistemas que sejam mais transparentes e responsáveis.

Relatório 1			
Nome do artigo:	AI models collapse when trained on recursively generated data	Referência:	SHUMAILOV, Ilia; SHUMAYLOV, Zakhar; ZHAO, Yiren; PAPERNOT, Nicolas; ANDERSON, Ross; GAL, Yarin. AI models collapse when trained on recursively generated data. *Nature*, v. 631, n. 8022, p. 755-759, 2024. Disponível em: https://www.nature.com/articles/s41586-024-07566-y. Acesso em: 10 fev. 2025.
Entendimento do Abstract:			
Com base no resumo lido, observou-se que o colapso do modelo é um fenômeno que pode ocorrer em qualquer modelo generativo. Esse colapso é caracterizado como um processo degenerativo no qual, em sua fase inicial, a distribuição original dos dados começa a perder as caudas, ou seja, os estados de menor probabilidade. Esse fenômeno tende a se tornar cada vez mais comum, pois o conteúdo disponível na internet vem sendo progressivamente influenciado por textos gerados por modelos, reduzindo a presença de conteúdo produzido exclusivamente por humanos. Como consequência, durante a raspagem de dados, a distribuição passa a refletir cada vez mais informações geradas artificialmente, acelerando ainda mais o colapso dos modelos.			

Entendimento do Discussion/Conclusão:
Diante da seção “Discussion” do artigo, foi comprovado que o colapso realmente acontece com modelos de IA generativa e, diante desse fato, é preciso garantir que os eventos de baixa probabilidade sejam considerados durante os reagrupamentos da distribuição de dados para o treinamento dos modelos, assegurando a geração de respostas justas ao usuário. Ademais, foram recomendadas algumas ações diante da problemática do colapso, como garantir que a distribuição original dos dados, proveniente da interação com conteúdo humano, faça parte do treinamento dos modelos sucessores. Além disso, sugeriu-se a criação de uma coordenação internacional que assegure o compartilhamento de informações entre diferentes grupos de desenvolvedores, com o objetivo de garantir o desenvolvimento de LLMs de qualidade, baseados em uma distribuição de dados confiável.

Relatório 2			
Nome do artigo:	Review of autonomous systems and collaborative AI agent frameworks	Referência:	JOSHI, Satyadhar. (2025). Review of autonomous systems and collaborative AI agent frameworks. International Journal of Science and Research Archive. 14. 961-972. 10.30574/ijsra.2025.14.2.0439. Disponível em: https://www.researchgate.net/publication/389068903_Review_of_autonomous_systems_and_collaborative_AI_agent_frameworks. Acesso em: 22 fev. 2025.
Entendimento do Abstract:			

O artigo em questão apresenta um panorama do uso atual dos agentes de IA, com foco em frameworks, destacando as principais ferramentas, suas vantagens e desvantagens. Além disso, explora a tecnologia dos agentes autônomos, enfatizando seu conceito, aplicações e aspectos técnicos, incluindo limitações e oportunidades. Também discute tendências futuras, oferecendo uma visão abrangente sobre o tema. O estudo segue com uma análise da aplicação dessa tecnologia em diferentes setores, como o financeiro, a gestão de riscos e o ambiente corporativo. Por fim, o artigo funciona como um guia, consolidando as principais observações recentes sobre a evolução e o impacto dos agentes autônomos.
Entendimento do Discussion/Conclusão:
A conclusão do presente artigo traz um panorama sobre o rápido avanço dos AI agents e as principais ferramentas para a construção dessa tecnologia, proporcionando uma visão geral de seu uso em diferentes cenários. Ademais, é enfatizada a aplicação dos AI agents em tarefas de alta complexidade. No entanto, o texto também destaca suas fragilidades e recomenda uma abordagem mais rigorosa em termos de ética e governança para essa tecnologia em ascensão.

Relatório 3			
Nome do artigo:	A Survey on Large Language Model based Autonomous Agents	Referência:	WANG, Lei; MA, Chen; FENG, Xueyang; ZHANG, Zeyu; YANG, Hao; ZHANG, Jingsen; CHEN, Zhiyuan; et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, v. 18, n. 6,

			2024, p. 186345. Disponível em: https://arxiv.org/abs/2308.11432. Acesso em: 22 de fev. 2025.
Entendimento do Abstract:			
O estudo em questão demonstra a evolução dos agentes autônomos com o auxílio de LLMs na realização de tarefas complexas. Ele explora o avanço desses agentes em atividades cada vez mais diversas em diferentes setores, oferecendo uma visão holística do tema. Além disso, apresenta as estratégias mais comuns para a integração dessas tecnologias e os desafios envolvidos.			
Entendimento do Discussion/Conclusão:			
A seção de “Conclusão” do artigo afirma que o estudo conduzido forneceu um panorama detalhado dos principais avanços de agentes auxiliados por LLMs, abordando sua construção, aplicação e evolução, além de citar aspectos técnicos. Por fim, a conclusão enfatiza que o artigo destaca os principais desafios e lacunas dessas ferramentas.			

Relatório 4			
Nome do artigo:	Can We Trust AI Agents? An Experimental Study Towards Trustworthy LLM-Based Multi-Agent Systems for AI Ethics	Referência:	**CERQUEIRA, José Antonio Siqueira de; et al.** Can we trust AI agents? An experimental study towards trustworthy LLM-based multi-agent systems for AI ethics. *arXiv preprint*, arXiv:2411.08881, 2024. Disponível em: https://arxiv.org/abs/2411.08881. Acesso em: 22 fev. 2025.

Entendimento do Abstract:
Neste estudo, foi analisado como os LLMs podem ajudar no desenvolvimento de IA ética. Foi criado um protótipo chamado LLM-BMAS, que usa múltiplos agentes para discutir questões éticas reais, gerando código ético e detalhado. O sistema abordou temas como viés, transparência, responsabilidade, consentimento e conformidade.
Entendimento do Discussion/Conclusão:
O estudo mostra técnicas para tornar os modelos de IA mais confiáveis na área de engenharia de software. Para isso, foi criado um sistema multi-agentes, onde cada um tinha uma função específica dentro do processo, ajudando a organizar as informações e melhorar a qualidade das respostas. Além disso, o sistema usou debates e conversas estruturadas para aprimorar a tomada de decisão.

Relatório 5			
Nome do artigo:	Transforming Competition into Collaboration: The Revolutionary Role of Multi-Agent Systems and Language Models in Modern Organizations	Referência:	CRUZ, Carlos Jose Xavier. Transforming competition into collaboration: The revolutionary role of multi-agent systems and language models in modern organizations. *arXiv preprint*, arXiv:2403.07769, 2024. Disponível em: https://arxiv.org/abs/2403.07769. Acesso em: 27 fev. 2025.
Entendimento do Abstract:			

Esse artigo fala sobre como combinar sistemas multiagentes (SMA) com grandes modelos de linguagem (LLM) pode mudar a forma como humanos interagem com agentes artificiais. A ideia é usar esses agentes para ajudar tanto em tarefas operacionais do dia a dia quanto em decisões estratégicas dentro das empresas. A abordagem do estudo propõe criar agentes baseados em LLM com perfis diferentes, que simulam comportamentos específicos e interagem entre si em um formato de conversa guiada.
Entendimento do Discussion/Conclusão:
O presente texto encerra enfatizando a interação entre multiagentes e LLMs, destacando seu alto impacto positivo em tarefas que exigem colaboração em organizações. Em seguida, apresenta um resumo das tarefas mais comumente realizadas com o uso de IA e conclui com uma reflexão sobre como essa interação tecnológica proporcionará novas formas de aplicação, reduzindo a complexidade e incentivando o uso criativo dessas tecnologias.

Relatório 6			
Nome do artigo:	Multi-Agent Large Language Models for Conversational Task-Solving	Referência:	BECKER, Jonas. Multi-agent large language models for conversational task-solving. *arXiv preprint*, arXiv:2410.22932, 2024. Disponível em: https://arxiv.org/abs/2410.22932. Acesso em: 22 fev. 2025.
Entendimento do Abstract:			
Este trabalho avalia sistemas multiagentes em tarefas conversacionais, analisando seu desempenho em diferentes paradigmas. Proponho uma taxonomia de 20 estudos			

(2022-2024) e um framework para LLMs multiagentes.
Entendimento do Discussion/Conclusão:
A presente conclusão aborda o principal tema do artigo, a interação entre multiagentes no contexto da comunicação. No entanto, o foco maior foi a relação entre os agentes na resolução de tarefas, destacando suas reações em diferentes cenários e o impacto da duração das conversas em sua performance. Além disso, ressalta-se que os agentes conseguem garantir a ética em suas interações, evitando temas inadequados. Por fim, o texto enfatiza que a introdução de LLMs no suporte aos multiagentes contribui significativamente para a resolução de tarefas complexas e para o alto desempenho.

Relatório 7			
Nome do artigo:	Strong Model Collapse	Referência:	DOHMATOB, Elvis; FENG, Yunzhen; SUBRAMONIAN, Arjun; KEMPE, Julia. Strong model collapse. *arXiv preprint*, arXiv:2410.04840, 2024. Disponível em: https://arxiv.org/abs/2410.04840. Acesso em: 27 fev. 2025.
Entendimento do Abstract:			
Este estudo analisa o colapso do modelo em redes neurais grandes, causado por dados sintéticos no treinamento. Mesmo 1% de dados sintéticos pode levar à degradação do desempenho, tornando inútil o aumento do conjunto de treinamento. Investiga-se também o impacto do aumento do tamanho do modelo, mostrando que modelos maiores podem amplificar o colapso.			

Relatório 8			
Nome do artigo:	How to Synthesize Text Data without Model Collapse?	Referência:	ZHU, Xuekai; CHENG, Daixuan; LI, Hengli; ZHANG, Kaiyan; HUA, Ermo; LV, Xingtai; DING, Ning; LIN, Zhouhan; ZHENG, Zilong; ZHOU, Bowen. How to Synthesize Text Data without Model Collapse? *arXiv preprint*, arXiv:2412.14689, 2024. Disponível em: https://arxiv.org/abs/2412.14689. Acesso em: 22 fev. 2025.
Entendimento do Abstract:			
O estudo analisa o impacto dos dados sintéticos no treinamento de modelos de linguagem, mostrando que uma maior proporção de dados sintéticos reduz o desempenho do modelo. Análises estatísticas indicam mudanças na distribuição dos dados e excesso de n-grams. Para evitar o colapso do modelo, propõe-se a edição de tokens em dados humanos para gerar dados semissintéticos. Experimentos confirmam que essa técnica melhora a qualidade dos dados e o desempenho do modelo.			

Entendimento do Discussion/Conclusão:
O artigo em questão conclui que a utilização de dados sintéticos pode comprometer a eficácia do pré-treinamento quando combinados com dados humanos, resultando em colapso não iterativo do modelo. Ademais, para mitigar esse problema, os autores propõem a edição em nível de token, adotando um método de reamostragem guiado por um modelo pré-treinado.

Relatório 9			
Nome do artigo:	The Impact of Large Language Models in Academia: from Writing to Speaking	Referência:	GENG, Mingmeng; CHEN, Caixi; WU, Yanru; CHEN, Dongping; WAN, Yao; ZHOU, Pan. The impact of large language models in academia: from writing to speaking. *arXiv preprint*, arXiv:2409.13686, 2024. Disponível em: https://arxiv.org/abs/2409.13686. Acesso em: 22 fev. 2025.
Entendimento do Abstract:			
O estudo mostra que os modelos de linguagem de grande porte (LLMs) estão impactando cada vez mais a sociedade humana, especialmente na informação textual. O impacto na fala está começando a surgir e tende a crescer no futuro, chamando a atenção para a influência implícita e o efeito cascata dos LLMs na sociedade humana.			
Entendimento do Discussion/Conclusão:			
O presente artigo aponta que, no contexto acadêmico, um número crescente de			

pessoas utiliza os padrões de respostas gerados por LLMs, influenciando tanto a escrita quanto a fala, especialmente a escrita. Consequentemente, o texto enfatiza o possível risco do colapso do modelo, considerando que, à medida que mais pessoas recorrem a essa ferramenta, inclusive na área acadêmica, aumenta a chance de obterem respostas de um modelo colapsado, ou seja, com vieses.

Relatório 10			
Nome do artigo:	Mitigating Social Bias in Large Language Models: A Multi-Objective Approach Within a Multi-Agent Framework	Referência:	XU, Zhenjie; CHEN, Wenqing; TANG, Yi; LI, Xuanying; HU, Cheng; CHU, Zhixuan; REN, Kui; ZHENG, Zibin; LU, Zhichao. Mitigating social bias in large language models: A multi-objective approach within a multi-agent framework. *arXiv preprint*, arXiv:2412.15504, 2024. Disponível em: https://arxiv.org/abs/2412.15504. Acesso em: 22 fev. 2025.
Entendimento do Abstract:			
Neste estudo, foi proposto uma abordagem multiobjetivo dentro de um framework multiagente (MOMA) para reduzir o viés social em LLMs sem prejudicar significativamente o desempenho. O MOMA utiliza múltiplos agentes para realizar intervenções causais nos conteúdos relacionados ao viés nas perguntas, quebrando a conexão direta entre esses conteúdos e as respostas.			
Entendimento do Discussion/Conclusão:			
A conclusão do artigo destaca as técnicas utilizadas para mitigar o viés dos modelos			

de LLMs, sendo uma das mais eficazes o uso de multi-agentes para abordar essa problemática.

Relatório 11			
Nome do artigo:	Towards Implicit Bias Detection and Mitigation in Multi-Agent LLM Interactions	Referência:	BORAH, Angana; MIHALCEA, Rada. Towards implicit bias detection and mitigation in multi-agent LLM interactions. *arXiv preprint*, arXiv:2410.02584, 2024. Disponível em: https://arxiv.org/abs/2410.02584. Acesso em: 22 fev. 2025.
Entendimento do Abstract:			
Neste estudo, modelos de LLMs estão sendo usados para obter insights sobre aspectos sociais, então é fundamental mitigar vieses. Neste artigo, foi investigado a presença de vieses implícitos de gênero em interações multiagente com LLMs e foi proposto duas estratégias para reduzi-los.			
Entendimento do Discussion/Conclusão:			
O presente estudo demonstra a presença de viés nos modelos de LLMs no contexto de gênero. Os pesquisadores desenvolveram técnicas de análise que evidenciaram a ocorrência de viés de forma recorrente. Além disso, diversas conclusões foram extraídas ao longo do estudo, sendo as principais: LLMs geram vieses mesmo quando treinados com dados produzidos por humanos; modelos de LLMs com maior quantidade de parâmetros apresentam maior propensão a vieses; a interação entre múltiplos agentes e LLMs pode agravar o viés; e o ajuste fino pode ser uma técnica eficaz para mitigar o viés no contexto de interação entre modelos de IA generativa e sistemas multiagentes.			

Relatório 12			
Nome do artigo:	Fairness in Multi-Agent AI: A Unified Framework for Ethical and Equitable Autonomous Systems	Referência:	RANJAN, Rajesh; GUPTA, Shailja; SINGH, Surya Narayan. Fairness in multi-agent AI: A unified framework for ethical and equitable autonomous systems. *arXiv preprint*, arXiv:2502.07254, 2025. Disponível em: https://arxiv.org/abs/2502.07254. Acesso em: 22 fev. 2025.
Entendimento do Abstract:			
Este artigo oferece uma visão abrangente sobre a equidade em IA multiagente, introduzindo um novo framework que integra restrições de equidade, estratégias de mitigação de vieses e mecanismos de incentivo para alinhar os comportamentos autônomos dos agentes com os valores sociais, equilibrando eficiência e robustez.			
Entendimento do Discussion/Conclusão:			
O presente artigo enfatiza o objetivo de criar um ambiente colaborativo entre pesquisadores para mitigar os vieses nas ações dos sistemas multiagentes, promovendo responsabilidade e transparência. Além disso, o estudo alerta para a necessidade de técnicas que minimizem o viés e garantam que os agentes atuem de maneira mais justa. Por fim, a pesquisa foi conduzida utilizando a modificação do sistema de recompensas como estratégia para mitigar ações indesejadas dos multiagentes.			

Relatório 13

Nome do artigo:	Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey	Referência:	ACHARYA, Deepak Bhaskar; KUPPAN, Karthigeyan; DIVYA, B. Agentic AI: Autonomous intelligence for complex goals – A comprehensive survey. *IEEE Access*, 2025. Disponível em: https://ieeexplore.ieee.org/document/10849561. Acesso em: 10 fev. 2025.

Entendimento do Abstract:
O estudo explora os conceitos fundamentais, as características únicas e as metodologias centrais que impulsionam o desenvolvimento dos agentes. Além disso, discute suas aplicações em áreas como saúde, finanças e software adaptativo, destacando as vantagens da implementação de sistemas agentes em cenários do mundo real. O estudo também aborda os desafios éticos relacionados a essa tecnologia, propondo soluções para questões como alinhamento de objetivos, restrições de recursos e adaptabilidade ao ambiente.
Entendimento do Discussion/Conclusão:
Na seção de “Conclusão”, o presente artigo destaca as diversas facetas dos agentes de IA, abordando seus conceitos, aplicabilidades e desafios. Além disso, enfatiza a ampla usabilidade desses sistemas em diversos cenários, mas também ressalta suas limitações. Por fim, alerta para a necessidade de uma governança mais robusta a fim de fortalecer a ética na aplicação dessa tecnologia.

Relatório 14

Nome do artigo:	Personalized Recommendation Systems using Multimodal, Autonomous, Multi Agent Systems	Referência:	THAKKAR, Param; YADAV, Anushka. *Personalized Recommendation Systems using Multimodal, Autonomous, Multi Agent Systems*. arXiv preprint arXiv:2410.19855, 2024. Disponível em: https://arxiv.org/abs/2410.19855. Acesso em: 22 fev. 2025.
Entendimento do Abstract:			
O artigo descreve um sistema de recomendação personalizado usando sistemas multimodais e multiagentes para melhorar a experiência do cliente no e-commerce. O sistema é composto por três agentes: o primeiro recomenda produtos, o segundo faz perguntas de acompanhamento com base em imagens e o terceiro realiza uma busca autônoma.			
Entendimento do Discussion/Conclusão:			
O estudo demonstra que houve colaboração entre os agentes em um sistema multiagente para auxiliar os usuários com recomendações de produtos. É importante destacar que a distribuição de dados utilizada ia além do histórico do cliente, incorporando também o uso de imagens.			

Relatório 15			
Nome do artigo:	Consumer reactions to technology in retail: choice	Referência:	ROHDEN, Simoni F.; ESPEARTEL, Lélis Balestrin. Consumer reactions to technology in retail: choice

	uncertainty and reduced perceived control in decisions assisted by recommendation agents		uncertainty and reduced perceived control in decisions assisted by recommendation agents. *Electronic Commerce Research*, v. 24, n. 2, p. 901-923, 2024. Disponível em: https://link.springer.com/article/10.1007/s10660-024-09808-7. Acesso em: 22 fev. 2025.
Entendimento do Abstract:			
A pesquisa destaca que agentes de recomendação podem reduzir a sobrecarga de escolhas e facilitar decisões de compra, mas também geram maior incerteza na tomada de decisão. Compras auxiliadas por esses agentes são percebidas como mais incertas, com menor controle percebido sobre as escolhas, resultando em menor satisfação e intenções de compra.			
Entendimento do Discussion/Conclusão:			
O presente estudo destaca os efeitos positivos do uso de agentes na redução da carga cognitiva do usuário durante a escolha e navegação de produtos. No entanto, os experimentos indicam que essa tecnologia pode aumentar a percepção de incerteza do usuário em relação às recomendações.			

Relatório 16			
Nome do artigo:	Investigating Bias in LLM-Based Bias Detection:	Referência:	LIN, Luyang; WANG, Lingzhi; GUO, Jinsong; WONG, Kam-Fai. Investigating bias in LLM-based bias detection: disparities

	Disparities between LLMs and Human Perception		between LLMs and human perception. *arXiv preprint*, arXiv:2403.14896, 2024. Disponível em: https://arxiv.org/abs/2403.14896. Acesso em: 22 fev. 2025.
Entendimento do Abstract:			
Nesta pesquisa, embora modelos de linguagem grandes (LLMs) robustos tenham surgido como ferramentas fundamentais para a previsão de viés, persistem preocupações sobre os vieses inerentes a esses modelos. Ademais, foi investigado a presença e a natureza do viés nos LLMs e seu impacto consequente na detecção de viés na mídia.			
Entendimento do Discussion/Conclusão:			
O texto enfatiza a presença de viés em modelos de LLMs e insiste na urgência de políticas, diretrizes e governança para mitigar essa problemática.			

Relatório 17			
Nome do artigo:	Bias of AI-generated content: an examination of news produced by large language models	Referência:	FANG, X.; CHE, S.; MAO, M. et al. Bias of AI-generated content: an examination of news produced by large language models. *Sci Rep*, v. 14, p. 5224, 2024. Disponível em: https://doi.org/10.1038/s41598-024-55686-2. Acesso em: 22 fev. 2025.

Entendimento do Abstract:
O estudo investiga o viés de gênero e racial no AIGC produzido por sete LLMs, incluindo ChatGPT e LLaMA, utilizando artigos de notícias do The New York Times e Reuters. A pesquisa revela que os LLMs demonstram vieses substanciais, especialmente contra mulheres e indivíduos da raça negra. O ChatGPT apresenta o menor nível de viés e é o único modelo capaz de recusar gerar conteúdo com prompts tendenciosos.
Entendimento do Discussion/Conclusão:
O texto enfatiza a presença de viés em modelos de LLMs, evidenciando que a AIGC (Conteúdo Gerado por IA) produzida por esses modelos apresenta vieses de gênero e raça em diferentes níveis. Destaca-se a eficácia do RLHF (Reforço a Partir de Feedback Humano) na mitigação desses vieses.

Relatório 18			
Nome do artigo:	A Multi-Agent Conversational Recommender System	Referência:	FANG, Jiabao; GAO, Shen; REN, Pengjie; CHEN, Xiuying; VERBERNE, Suzan; REN, Zhaochun. A multi-agent conversational recommender system. *arXiv preprint*, arXiv:2402.01135, 2024. Disponível em: https://arxiv.org/abs/2402.01135. Acesso em: 22 fev. 2025.
Entendimento do Abstract:			

O artigo propõe o Sistema de Recomendação Conversacional Multi-Agente (MACRS), que melhora o fluxo de diálogo e coleta de preferências do usuário. O MACRS usa uma estrutura cooperativa de múltiplos agentes para gerar e escolher respostas adequadas e um mecanismo de reflexão para ajustar o planejamento do diálogo com base no feedback do usuário.
Entendimento do Discussion/Conclusão:
O estudo demonstra as técnicas utilizadas para aprimorar a abordagem de recomendação ao usuário, adotando um sistema de multiagentes, no qual cada agente é responsável por uma parte da estratégia de diálogo, com o suporte de modelos de LLMs. Além disso, foi empregado um mecanismo de feedback contínuo do usuário e a integração de suas informações para aumentar a precisão dos agentes.

Relatório 19			
Nome do artigo:	Poisoning and Backdooring Contrastive Learning	Referência:	CARLINI, Nicholas; TERZIS, Andreas. Poisoning and backdooring contrastive learning. *arXiv preprint*, arXiv:2106.09667, 2021. Disponível em: https://arxiv.org/abs/2106.09667. Acesso em: 10 fev. 2025.
Entendimento do Abstract:			
Neste estudo, mesmo envenenando apenas 0,01% de um conjunto de dados, foi mostrado que é possível induzir o modelo a cometer erros, levantando questões sobre a viabilidade de treinar com dados não curados da internet.			
Entendimento do Discussion/Conclusão:			

O estudo mostra como o uso de conjuntos de dados não filtrados pode aumentar os riscos de ataques de envenenamento em modelos de aprendizado de máquina. Ele explica que modelos modernos treinam com grandes volumes de dados retirados da Internet, sem uma revisão rigorosa, o que facilita a inserção de informações maliciosas por adversários. Os pesquisadores demonstraram que esses ataques podem ser feitos com menos esforço do que em métodos tradicionais e que aumentar a quantidade de dados não impede os ataques. Para resolver esse problema, o estudo sugere que novas formas de defesa sejam desenvolvidas, já que revisar manualmente todos os dados não é viável.