Jones Martimiano Coelho

O Impacto Da Qualidade De Dados Em Grandes Modelos De Linguagem: Uma Revisão De Escopo Da Literatura

São Paulo
2024

Jones Martimiano Coelho

O Impacto Da Qualidade De Dados Em Grandes Modelos De Linguagem: Uma Revisão De Escopo Da Literatura

Trabalho apresentado à Escola Politécnica da Universidade de São Paulo para obtenção do Certificado de Especialista em Engenharia de Dados e Big Data.

Orientador:

Wesley Lourenço Barbosa

São Paulo
2024

Catalogação-na-publicação

Coelho, Jones Martimiano
O Impacto da Qualidade de Dados em Grandes Modelos de Linguagem: Uma Revisão de Escopo da Literatura / J. M. Coelho -- São Paulo, 2024.
53 p.

Monografia (Especialização em Engenharia de Dados e Big Data) - Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia.

1.qualidade de dados 2.grandes modelos de linguagem 3.revisão da literatura 4.inteligência artificial I.Universidade de São Paulo. Escola Politécnica. PECE – Programa de Educação Continuada em Engenharia II.t.

AGRADECIMENTOS

Agradeço primeiramente à minha família, pelo amor incondicional, paciência e incentivo ao longo de toda essa jornada. Sem o apoio de vocês, este momento não seria possível. Aos meus professores, agradeço pela dedicação, pelos ensinamentos valiosos e pelo constante estímulo ao conhecimento, que me permitiram alcançar este objetivo. E, finalmente, aos meus amigos, que estiveram ao meu lado nos momentos de desafio, oferecendo palavras de encorajamento e apoio, o meu sincero agradecimento.

Este trabalho é fruto do suporte e confiança de todos vocês.

"A qualidade dos dados determina a qualidade da tomada de decisões. Em um mundo de conjuntos de dados cada vez mais complexos, garantir a integridade dos dados é a chave para desbloquear insights e inovações verdadeiras."

-- Thomas Redman

RESUMO

Este estudo apresenta uma revisão de escopo da literatura sobre o impacto da qualidade dos dados no desenvolvimento de grandes modelos de linguagem (LLMs) e inteligências artificiais generativas. A pesquisa aborda questões relacionadas ao desempenho, confiabilidade e vieses, destacando como a qualidade de dados afeta diretamente a generalização e precisão desses modelos. Os métodos incluem a análise de artigos publicados entre 2018 e 2024, identificando desafios, soluções e práticas para melhorar a curadoria e o gerenciamento de dados. Foram analisados 39 trabalhos e os resultados indicam que práticas como o pré-processamento de dados e *frameworks* de governança podem mitigar vieses, reduzir erros e aumentar a eficiência. O estudo enfatiza que a qualidade dos dados é um fator crítico para o sucesso dos LLMs, influenciando desde o desempenho técnico até implicações éticas e sociais.

Palavras-Chave – Qualidade de Dados, Grandes Modelos de Linguagem, Inteligência Artificial Generativa, Vieses, Governança de Dados.

ABSTRACT

This study presents a scope review of the literature on the impact of data quality on the development of large language models (LLMs) and generative artificial intelligence. The research addresses issues related to performance, reliability, and biases, emphasizing how data quality directly affects the generalization and accuracy of these models. The methodology includes the analysis of articles published between 2018 and 2024, identifying challenges, solutions, and practices for improving data curation and management. A total of 39 studies were analyzed, and the results indicate that practices such as data preprocessing and governance frameworks can mitigate biases, reduce errors, and increase efficiency. The study underscores that data quality is a critical factor for the success of LLMs, influencing not only technical performance but also ethical and social implications.

Keywords – Data Quality, Large Language Models, Generative Artificial Intelligence, Biases, Data Governance.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BERT Bidirectional Encoder Representations from Transformers. 12, 13, 18, 19

ELMo Embeddings from Language Models. 18

GPT *Generative Pre-trained Transformer*. 12, 13, 15, 18, 19, 22

IA Inteligência Artificial. 13, 15, 17, 21, 25, 39–43

IFD Instruction-Following Difficulty. 34

LLM *Large Language Model*. , 12–15, 17–26, 28–36, 38–41, 43

NLP *Natual Language Processing*. 13

SUMÁRIO

1 INTRODUCÃO

Nas primeiras décadas do século XXI observou-se um aumento significativo no desenvolvimento de novas tecnologias, especialmente nas áreas de telecomunicações e computação. A informatização de diversos setores da economia tornou ubíqua a presença de dispositivos eletrônicos, levando ao surgimento de uma onda sem precedentes de acúmulo de dados (SCHWAB, 2017; NAIMI; WESTREICH, 2014). O barateamento das tecnologias de armazenamento e o aumento da capacidade de processamento permitiram que as organizações acumulassem vastas quantidades de informações. Esse fenômeno impulsionou o reconhecimento do valor estratégico dos dados, tornando a capacidade de coletar, armazenar e, principalmente, analisar grandes volumes de dados essencial para empresas que buscam manter-se competitivas em um mercado dinâmico e orientado por dados. Esse processo de acumulação e processamento de grandes quantidades de dados foi denominado *Big Data*. Segundo Chen, Chiang e Storey (2012), o conceito de *Big Data* é tradicionalmente descrito pelos três "Vs": Volume, Velocidade e Variedade, que se referem, respectivamente, à grande quantidade de dados gerados, à rapidez com que são processados e à diversidade de formatos e fontes de dados. Com o passar do tempo, outros "Vs"foram adicionados, como Veracidade, que diz respeito à qualidade dos dados, e Valor, relacionado à utilidade dos dados, destacando a crescente complexidade da análise no contexto de *Big Data* (CHEN; CHIANG; STOREY, 2012).

Com o aumento da disponibilidade de conteúdo digital e o crescimento exponencial dos dados em diversos formatos, de publicações em redes sociais, páginas na internet a publicações de artigos acadêmicos e entre outros, foi possível criar um grande conjunto de dados brutos para treinar grandes modelos de linguagem, do inglês *Large Language Model* (LLM). Esse célere aumento permitiu que modelos como *Generative Pre-trained Transformer* (GPT) e Bidirectional Encoder Representations from Transformers (BERT) conseguissem ter acesso a mais dados em seus conjuntos de treinamento, o que tem sido fundamental para o desenvolvimento das atuais capacidades, tanto em desempenho quanto em generalização (DEVLIN et al., 2019; RADFORD et al., 2019). A proliferação de conjuntos de dados massivos, diversos e heterogêneos tem sido crucial para as habilidades

dos LLM de executar tarefas de aprendizado, melhorando assim de forma significativa a adaptação e entendimento contextual em diferentes domínios (BROWN et al., 2020).

Assim, o advento de LLM e inteligências artificiais generativas tem revolucionado o campo de *Natual Language Processing* (NLP) (ou Processamento de Linguagem Natural), permitindo avanços em tarefas como tradução automatizada, criação de conteúdo, geração de código e *ChatBots*. Modelos como GPT-3, *ChatGPT*, *Llamma 2*, e *Code Llama* que têm demonstrado capacidades notáveis em entender e gerar texto imitando as capacidades humanas, têm sido amplamente usados em diferentes indústrias. Porém, o desempenho e a confiabilidade destes modelos, assim como modelos de aprendizado de máquina tradicional, estão intrinsecamente ligados à qualidade dos dados de treinamento. Portanto, qualidade de dados se mostra, novamente, como um fator crítico não apenas para a eficiência de LLM, mas também para tendências que esses modelos têm de gerarem resultados enviesados e/ou errados.

A qualidade de dados usados no treinamento impacta o desempenho por afetarem diretamente a capacidade do LLM de generalizar e produzir resultados confiáveis. Dados com baixa qualidade podem introduzir vieses, erros e limitações à capacidade de generalização do modelo (MYERS et al., 2023). A homogenização de modelos base como o GPT ou BERT amplificam quaisquer vieses presentes nos dados de treinamento, afetando todas as aplicações e casos de uso. Os desafios relacionados à qualidade de dados no treinamento de LLMs são diversos. Um dos principais desafios é garantir a diversidade e representação nas bases de treinamento, enquanto minimiza vieses herdados de bases de dados extraídas da internet. Penedo et al. (2023) discutem o custo computacional de analisar grandes conjuntos de dados e o risco de introduzir vieses ao realizar filtros em excesso ou mesmo remover informações importantes para o treinamento do modelo. Somado a isso, a escassez de dados e a variabilidade nas fontes de dados de treinamento impactam o viés e a confiabilidade do modelo, como destacado no desenvolvimento dos modelos *Llama 2* (TOUVRON et al., 2023).

Garrido-Muñoz, Martínez-Santiago e Montejo-Ráez (2023) evidenciam como vieses de gênero herdados de dados de treinamento levam a resultados distorcidos e com potencial de serem prejudiciais em LLM, focando na necessidade de uma curadoria decente e estratégias para minimizar vieses. Além disso, Schwabe et al. (2024) pontuam que má qualidade de dados resulta em vieses e precisões baseadas em IA não acuradas, que no contexto médico, podem impactar a evolução dos pacientes ao propagar vieses, imprecisões e inconsistências nas previsões clínicas, podendo levar a diagnósticos, tratamentos e alocações de recursos incorretos.

Para resolver esses problemas, diversas soluções têm sido propostas nos últimos anos. Liu et al. (2024) apresentaram o *CoachLM*, um modelo projetado para melhorar a qualidade de dados através da revisão automatizada de pares de instruções de baixa qualidade. Essa abordagem utiliza dados analisados por especialistas para treinar um LLM em menor escala, chamado *CoachLM*, que, por sua vez, revisa as bases de dados usadas para realizar o refinamento de um LLM alvo, melhorando assim o desempenho, sem descartar informações relevantes. Na mesma abordagem, Jain et al. (2024) demonstram que aumentar a qualidade de dados através de transformações estruturadas de código pode aumentar o desempenho de modelos que geram código, tendo casos onde os modelos que passaram pelo processo de refinamento em dados mais limpos têm um aumento no desempenho de até 30% quando comparado a modelos que foram treinados em dados brutos.

Também, Li et al. (2024a) propõem um método que aumenta a qualidade de instruções de refinamento através de colaboração entre modelos "professores"e modelos "alunos". Essa abordagem permite que o modelo "aluno"incorpore, de forma seletiva, pares de "instruções-resposta"refinadas, melhorando assim a compatibilidade e reduzindo o viés. Bojic et al. (2023) apresenta um *framework* centralizado em dados para melhorar bases de dados especializadas para o treinamento de modelos voltados à compreensão de leitura de textos. No trabalho é enfatizado que filtrar os dados de treinamento e aumentar a base de dados resultam em um aumento no desempenho dos modelos.

Um *framework* para limpeza de dados interativa usando um LLM para detectar e reparar erros é proposto por Ni et al. (2024). A ferramenta, chamada *IterClean*, mostra significativos avanços em qualidade de dados e desempenho dos modelos. Qian, Reif e Kahng (2024) elencam os desafios enfrentados por profissionais que atuam com LLMs em uma grande empresa de tecnologia por causa da falta de definições padronizadas e métodos de validação para qualidade de dados no contexto de LLMs. No trabalho, profissionais entrevistados realçam que muitas das vezes se baseiam em intuição e ferramentas pontuais. Isso ressalta a necessidade de uma avaliação sistemática da qualidade dos dados e de *frameworks* padronizados.

1.1 Objetivo

Este trabalho tem como objetivo avaliar o impacto da qualidade de dados no desenvolvimento de LLMs e inteligências artificiais generativas.

1.2 Justificativa

Essa pesquisa se justifica devido ao papel crítico que a qualidade de dados tem no desenvolvimento e implantação de LLM e IA Generativa. Como IA tem cada vez mais integrado diversos setores da sociedade, garantir que esses sistemas produzam resultados acurados, sem vieses e confiáveis é de suma importância. Ao endereçar os desafios associados com a qualidade de dados e explorar as soluções, será possível melhorar o desempenho de LLMs e mitigar os riscos.

1.3 Metodologia

A metodologia adotada neste trabalho é predominantemente sistemática, com foco na identificação, análise e interpretação de estudos primários relevantes para as questões de pesquisa. A abordagem empregada garante a reprodutibilidade e minimiza vieses.

Para estruturar a revisão, as três fases principais descritas por Kofod-Petersen (2015) e Kitchenham (2007) foram seguidas: planejamento, execução e síntese. O planejamento consiste na formulação das perguntas de pesquisa, desenvolvimento do protocolo e definição dos critérios de inclusão e exclusão. A execução envolve a busca e seleção dos trabalhos, enquanto a síntese trata da análise e integração dos dados extraídos.

No planejamento, foram definidas as perguntas principais, descritas no capítulo 3. O protocolo documenta o processo, incluindo fontes a serem consultadas, e palavras-chave relacionadas ao tema.

Durante a execução, foram aplicados os critérios de inclusão e exclusão. Para apoiar na seleção inicial, foi utilizado o *ChatGPT*-4 para classificar os trabalhos encontrados de acordo com o nível de relevância quando considerado os critérios. Estudos duplicados ou de qualidade insuficiente foram excluídos, garantindo a integridade dos trabalhos selecionados.

Na fase de síntese, foi empregado uma análise qualitativa, que permite a identificação de padrões e tendências gerais. A estruturação final dos dados segue a recomendação de Kitchenham (2007), organizando os resultados em tabelas e gráficos a fim de prover uma melhor visualização e compreensão.

2 CONTEXTUALIZAÇÃO TEÓRICA

2.1 Qualidade de Dados: Uma breve história

A disciplina de qualidade de dados se desenvolveu muito ao longo das últimas décadas. Essas mudanças refletem as evoluções que aconteceram nas disciplinas de gestão de dados, capacidade computacional e aplicações de dados. Durante as décadas de 1970 e 1980, os esforços para a qualidade de dados concentraram-se principalmente em sistemas de gerenciamento de bancos de dados, com foco em garantir a precisão, consistência e completude das bases de dados dentro das organizações. À medida que o conceito de bancos de dados relacionais amadurecia, os pesquisadores ampliaram o escopo da qualidade de dados para abordar a minimização de redundâncias, restrições de integridade e consultas a dados estruturados (BATINI; LENZERINI; NAVATHE, 1986). Na década de 1980, a compreensão sobre qualidade de dados evoluiu ainda mais, incorporando dimensões como pontualidade, interpretabilidade e acessibilidade, enfatizando o alinhamento das estruturas de banco de dados com as necessidades organizacionais e fundamentos ontológicos (WAND; WANG, 1996). Os primeiros *frameworks* focavam na integridade dos dados e na normalização como aspectos fundamentais da gestão da qualidade. Por exemplo, o Modelo Relacional de Dados de Codd introduziu conceitos de normalização, como a Primeira Forma Normal (1NF) e restrições de integridade para reduzir redundâncias e garantir a consistência dos dados (CODD, 1970). Isso foi posteriormente refinado com a Forma Normal de Boyce-Codd (BCNF), que abordava anomalias não resolvidas por formas normais anteriores. Além disso, o *Schema Integration Framework* enfatizava a consistência e a completude durante o design de esquemas (BATINI; LENZERINI; NAVATHE, 1986). Mais tarde, o *Ontological Data Quality Framework* vinculou dimensões de qualidade de dados, como integridade e consistência, à semântica organizacional (WAND; WANG, 1996).

Durante a decada de 1990, com o surgimento dos *Data Werehouses* e *Business Intelligence*, o tema de qualidade de dados ganhou mais atenção, pois as organizações passaram a consolidar dados de diferentes fontes. O foco mudou de integridade para limpeza de

dados (*Data Cleaning*) e técnicas de deduplicação, como discutido nos trabalhos pioneiros de Redman (1998), no qual é enfatizado o custo econômico de qualidade de dados ruim e propõe soluções sistemáticas para resolver. Governança de dados e *frameworks* de administração de dados surgiram, enfatizando a responsabilidade organizacional e o gerenciamento de dados através de sistemas (ENGLISH, 1999).

Os anos 2000 viram o surgimento da internet e dados digitais, o que trouxe novas oportunidades e desafios. A explosão de dados gerados pelas plataformas online criou a demanda por ferramentas de qualidade robustas para lidar com a natureza diversa e não estruturada dos dados digitais. Durante essa fase, as dimensões de qualidade de dados passaram a englobar pontualidade, acessibilidade e a relevância, conforme descrito por Wang e Strong (1996). Avanços em sumarizações, integração de dados e processamento em tempo real marcaram pontos importantes da disciplina, refletindo o aumento da complexidade nos ecossistemas de dados (BLEIHOLDER; NAUMANN, 2009; BRUCKNER; LIST; SCHIEFER, 2002).

Na década passada, o advento de *Big Data* e Inteligência Artificial (IA) demandaram uma reformulação de como a questão de qualidade de dados é abordada. Heterogeneidade, volume e velocidade passaram a ser os novos problemas que necessitavam de soluções escaláveis de qualidade de dados. Pesquisas passaram a enfatizar a qualidade como "adequação ao uso"(*fitness for purpose*), conforme descrito em Zhu et al. (2014), convergindo características dos dados com as necessidades de análise de dados e modelos de inteligência artificial. Neste período, as pesquisas em qualidade de dados passam a se relacionar com ética, redução de vieses, *compliance*, por causa dos impactos crescentes de IA na sociedade.

Atualmente, a qualidade de dados continua sendo um fator crucial para aprendizado de máquina e aplicações de inteligência artificial, onde decisões baseadas em dados dependem bastante da integridade dos dados. Conforme LLMs e outras técnicas de IA se espalham, pesquisas em qualidade de dados têm sido expandidas para lidar com os desafios únicos ao lidar com viés de modelos, processamento de dados não estruturados (como imagem, som, publicações em redes sociais e entre outros) e escalabilidade das soluções (ZHANG; ABDUL-MAGEED; LAKSHMANAN, 2024; LAKRETZ et al., 2022).

2.2 Grandes Modelos de Linguagem

LLM é uma rede neural treinada de última geração, projetada para processar e gerar texto de forma semelhante a humanos. Essa classe de modelos é construída baseada na arquitetura *Transformer*, apresentada por Vaswani et al. (2017), na qual são utilizados mecanismos de autoatenção (*self-attention*) para lidar com dependências em sequências de dados de forma eficiente. Ao modelar simultaneamente as relações em toda a sequência, os *Transformers* superam arquiteturas anteriores, como redes recorrentes, na captura de padrões complexos da linguagem, proporcionando maior eficiência e precisão em tarefas como tradução automática e compreensão de texto (ZHANG; ABDUL-MAGEED; LAKSHMANAN, 2024; LAKRETZ et al., 2022; KUMAR, 2024).

A evolução dos LLMs começou com pesquisas em *word embeddings*, isto é, representações vetoriais densas de palavras, projetadas para capturar seu significado semântico com base no contexto em que aparecem, como é o caso do *Word2Vector*, que representava palavras como vetores com dimensões fixas (MIKOLOV, 2013). Avanços no tema introduziram *embeddings* contextuais com modelos como Embeddings from Language Models (ELMo) e BERT, que conseguiam capturar o significado das palavras dependendo do contexto (PETERS et al., 2018; DEVLIN et al., 2019). A introdução de *Transformers* no GPT-1 em 2018 marcou um ponto de virada, o que levou à criação de modelos maiores e mais capazes como GPT-3 e GPT-4, assim como soluções abertas como o *BLOOM* e o *LLaMa* (BROWN et al., 2020; TOUVRON et al., 2023).

LLMs são pré-treinados em grandes volumes de dados utilizando aprendizado não supervisionado, o que permite que os modelos identifiquem padrões linguísticos fundamentais que vão além de contextos e domínios específicos. Essa capacidade resulta na geração de representações universais de linguagem, ou seja, abstrações que codificam características semânticas e sintáticas da língua de maneira adaptativa e com um bom desempenho. Esses modelos pré-treinados são então ajustados para aplicações específicas, como resumo de texto ou tradução (MYERS et al., 2023). O princípio da escalabilidade, isto é, a ideia de que aumentar a quantidade de parâmetros dos modelos e o volume de dados de treinamento, melhora o desempenho, tem levado ao desenvolvimento de modelos como GPT-3 (com 175 bilhões de parâmetros) e o *LLaMA2* (com até 70 bilhões de parâmetros), demonstrando a efetividade da "Lei da Escala"(*scaling laws*) no desenvolvimento de LLMs (TOUVRON et al., 2023; KAPLAN et al., 2020). Na Tabela 1 temos uma comparação entre os maiores modelos de linguagem.

Tabela 1: Comparação entre os maiores LLMs

Modelo	**Ano**	**Quantidade de Parâmetros**	**Arquitetura**	**Principais Características**
GPT-3	2020	175 bilhões	*Transformer*	Aprendizado com poucos exemplos, escalabilidade
BERT	2019	340 milhões	*Transformer*	Significado das palavras baseado no contexto, bidirecionalidade
LLaMA 2	2023	7-70 bilhões	*Transformer*	Código aberto, específico para modelos de *chat*
GPT-4	2023	>175 bilhões	*Transformer*	diversas aplicações, raciocínio avançado

Fonte: Autor

LLMs têm um potencial disruptivo em diversos setores, porém, também apresentam desafios sociais e éticos. O treinamento em grandes bases de dados, principalmente quando o conteúdo é extraído da internet, existe o risco de incorporar na memória do modelo informações sensíveis e privadas, o que apresenta um problema para a privacidade de dados. Dado isso, é necessário implantar protocolos rígidos de limpeza de dados (MYERS et al., 2023). Também relacionado aos dados de treinamento, qualquer viés contido nos dados pode resultar em *outputs* discriminatórios, afetando aplicações em domínios sensíveis, como é o caso de saúde e legal (TOUVRON et al., 2023). Por fim, devido à natureza probabilística dos LLMs, é possível que as respostas geradas soem plausíveis, porém são imprecisas ou infundadas (KUMAR, 2024). Abordar esses desafios envolve diretrizes éticas, transparência do modelo e avaliação contínua para alinhar as saídas dos LLMs com os valores da sociedade.

3 REVISÃO DE ESCOPO DA LITERATURA

A abordagem para esta revisão de escopo foi exploratória e descritiva, a fim de sintetizar o conhecimento atual e identificar temas, metodologias e desafios-chave relacionados à qualidade dos dados no desenvolvimento de LLMs. Uma abordagem exploratória foi apropriada, dado o caráter em rápida evolução das tecnologias de LLMs e a gama de questões de qualidade de dados que elas abrangem. Ao explorar diversas perspectivas e inovações nas práticas de qualidade de dados, esta revisão buscou revelar tendências, destacando melhores práticas e identificando lacunas na literatura existente. Todos os dados e informações foram obtidos de fontes públicas e, quando não, foi solicitado diretamente aos autores do estudo, com as devidas citações aos trabalhos originais em respeito à propriedade intelectual e integridade acadêmica.

A condução da revisão foi estruturada seguindo a metodologia sistemática proposta pela plataforma *Parsif.al* [1], a qual estabelece um conjunto de etapas bem definidas para garantir a abrangência e a rigorosidade do processo baseado na proposta de Kitchenham (2007).

O protocolo utilizado contém três fases e cada fase é composta por alguns subitens:

- **Planejamento**
  - Definição do objetivo da revisão.
  - Definição do PICOC.
  - Escolha das perguntas de pesquisa.
  - Definição das palavras-chave e seus respectivos sinônimos.
  - Construção da *string* de busca.
  - Escolha das fontes de dados.
  - Definição dos critérios de inclusão e exclusão.

[1] ⟨https://parsif.al/about/⟩

  - Definição dos critérios de avaliação da qualidade dos artigos.

- **Condução**
  - Busca dos artigos.
  - Seleção dos artigos baseados nos critérios de inclusão e exclusão.
  - Atribuição dos critérios de qualidade.
  - Extração de dados dos artigos.

- **Resultados**

3.1 Objetivo

Apresentado na subseção 1.1.

3.2 PICOC

O *framework* PICOC, conforme delineado em metodologias de revisão sistemática, é uma abordagem estruturada projetada para formular perguntas de pesquisa de forma abrangente. PICOC é um acrônimo para *Population* (População), *Intervention* (Intervenção), *Comparison* (Comparação), *Outcome* (Resultado) e *Context* (Contexto), fornecendo uma estrutura clara para orientar revisões sistemáticas. Esse *framework* garante que todos os aspectos de uma pergunta de pesquisa sejam explicitamente abordados, auxiliando no desenvolvimento de estratégias de busca precisas e na avaliação da relevância dos estudos. O PICOC facilita uma abordagem rigorosa e replicável para sintetizar evidências, garantindo que as descobertas sejam tanto relevantes quanto acionáveis Kitchenham (2007).

A partir da definição do objetivo apresentado na subseção 1.1, a População a ser analisada são os grandes modelos de linguagem. Como busca-se entender como a qualidade de dados impacta no desenvolvimento de LLMs, qualidade de dados é a Intervenção. Neste trabalho não busca-se comparar soluções e/ou *frameworks*, sendo assim o item Comparação não se aplica. O Resultado é o impacto de qualidade de dados no desenvolvimento de LLMs, analisado aqui qualitativamente. E por fim, como não definiu-se nenhuma aplicação específica de IA, o Contexto não se aplica.

3.3 Perguntas de Pesquisa

Como a pesquisa tem o intuito de reunir evidências e conhecimento de como a disciplina de qualidade de dados está sendo aplicada no desenvolvimento de LLMs, foram definidas as perguntas de pesquisa apresentadas na Tabela 2. As perguntas são importantes para guiar a extração de dados realizada nos artigos selecionados.

Tabela 2: Perguntas de pesquisa

	Perguntas de Pesquisa
1	Como a qualidade de dados impacta o desempenho e confiança de LLMs?
2	Quais são os desafios e soluções existentes para a qualidade de dados no treinamento de LLMs?
3	Como os problemas de qualidade de dados contribuem para vieses e erros em inteligencias artificiais generativas?

Fonte: Autor

3.4 Fontes de Dados e Estratégia de Busca

A fim de obter artigos e conteúdo relevante para esta revisão, foram selecionados quatro repositórios de publicações. Os repositórios selecionados foram escolhidos devido à reputação, foco no tema e acervo. Os repositórios são: *IEEE Explore*, *Scopus*, *ACM Digital Library* e *SpringerLink*

Com o objetivo de utilizar os mecanismos de busca destes repositório, uma *string* de busca foi criada. Para a construção da *string* foram utilizadas uma combinação das seguintes palavras e siglas: LLM, *Large Language Model*, *Generative AI*, GPT e *data quality*. A utilização de termos em inglês é consequência de ser a língua mais utilizada nas publicações nos repositórios escolhidos para as buscas. Para garantir uma busca ótima, foram utilizados operadores lógicos da busca avançada das plataformas de busca de artigos. Assim, a *string* utilizada foi a definida na Tabela 3.

Tabela 3: *String* de busca

STRING de Busca
`("LLM" OR "Large Language Model" OR "Generative AI" OR "GPT") AND "data quality"`

Fonte: Autor

Com o intuito de reduzir a quantidade de artigos a serem analisados, foram definidos critérios de exclusão e inclusão. Os critérios de inclusão e exclusão estão apresentados na Tabela 4.

Tabela 4: Critérios de inclusão e exclusão

	Inclusão	Exclusão
1	Artigos publicados desde o dia 01 de janeiro de 2018 até 19 de outubro de 2024, tendo assim um período de quase seis anos de publicações.	Artigo aborda a aplicação de LLMs para resolver qualidade de dados, mas não discute qualidade de dados para LLMs em si.
2	Artigos devem ter sido escritos em inglês.	
3	Deve ser um artigo cientifico.	
4	Artigo deve abordar qualidade de dados no contexto de LLMs.	
5	Artigo acessível de forma gratuita, seja por acesso público ou através de parceria com a Universidade de São Paulo.	

Fonte: Autor

A *string* de busca foi executada em cada repositório de artigos, aplicando-se filtros de data de publicação, idioma e tipo de documento, resultando em 982 resultados no total. Os metadados dos artigos foram então baixados para o computador no formato *BibTeX*. Optou-se por esse formato porque o *BibTeX* é uma ferramenta eficiente de gerenciamento de referências que permite formatar dados bibliográficos em documentos LaTeX, proporcionando estilos de citação padronizados, facilidade de atualização e integração otimizada para a escrita acadêmica (PATASHNIK, Oren, 1988). Então, os arquivos *BibTex* foram importados na ferramenta *Mendeley*. O *Mendeley* é uma ferramenta de gerenciamento de referências e uma rede social acadêmica que auxilia os pesquisadores a organizar, anotar e compartilhar artigos de pesquisa, além de permitir a integração perfeita de citações durante a escrita (Mendeley, 2024).

Uma vez os artigos importados para o *Mendeley*, o passo seguinte foi remover os artigos duplicados. Nesta etapa, foram removidos 22 trabalhos.

Dado o volume de 961 artigos únicos a serem analisados, foi utilizada a ferramenta *ChatGPT* na sua versão 4.0 para aplicar o critério de inclusão 4 da Tabela 4. O *ChatGPT-4.0* tem a capacidade de gerar escores de similaridade entre sequências de texto ao utilizar *embeddings* gerados a partir de sua compreensão contextual da linguagem. Isso permite que ele compare textos com base na relevância semântica dentro de um determinado

contexto, tornando-o particularmente eficaz para tarefas como detecção de paráfrases, avaliações de similaridade contextual e recuperação de informações.

Para aproveitar as capacidades da ferramenta, foi criado um comando, no qual foi providenciado o resumo e o título do artigo, e solicitou-se para que o LLM gerasse uma pontuação entre 0 e 100. Essa pontuação foi utilizada como uma medida para classificar o quão correlacionado o artigo está com o tema e as perguntas de pesquisa. O comando utilizado é mostrado na Figura 1. O comando foi enviado através da API da OpenAI para interação com o *ChatGPT-4.0*. O código completo pode ser acessado neste repositório do *GitHub*[2].

Figura 1: Comando LLM

```
prompt = """
Analyze the following papers and provide a relevance score (0-100) for each one
↪ based on their usefulness for a literature review to understand the
↪ importance of data quality for LLMs and Generative AI..
Please evaluate the papers with respect to the following research questions:
1. How does data quality impact the performance and reliability of LLMs?
2. What are the current challenges and solutions related to data quality in
↪ training LLMs?
3. How do data quality issues contribute to biases or errors in Generative AI
↪ outputs?

Format the output as a JSON object with the following structure:
'[{"Title": "[title]", "Score": "[score]", "Reason": "[brief explanation]"}]'
Here are the papers:
"""
```

Fonte: Autor

O código gerou um arquivo de texto que contém o nome do artigo, sua pontuação e uma breve explicação sobre como o artigo se relaciona com o tema. A Tabela 5 apresenta uma amostra de cinco artigos com suas respectivas pontuações. A listagem completa está disponível no repositório do GitHub[3].

Utilizando uma ferramenta de gerenciamento de planilhas, foi realizada uma análise preliminar das pontuações e títulos dos artigos. Além disso, por meio de amostragem, validou-se o conteúdo dos resumos dos artigos para verificar se as pontuações atribuídas

[2] ⟨https://github.com/Jones-Coelho/literature-review-data-quality-llm/blob/master/code/filtro_artigos_llm.ipynb⟩

[3] ⟨https://github.com/Jones-Coelho/literature-review-data-quality-llm/blob/master/code/resultado_analise_llm.json⟩

Tabela 5: Amostra da pontuação dos artigos pelo *ChatGPT-4.0*

Título	Pontuação	Explicação
(ZHOU et al., 2024b)	95	Este artigo fornece uma visão abrangente das dimensões e ferramentas de qualidade de dados, abordando diretamente o impacto da qualidade dos dados em modelos de ML, incluindo LLMs. Discute desafios e soluções, tornando-o altamente relevante para todas as três perguntas de pesquisa.
(LI et al., 2024b)	80	Foca na melhoria da qualidade dos dados para ajuste de instrução de LLMs, abordando desafios e soluções relacionados à qualidade dos dados em LLMs.
(HAGER et al., 2024)	75	Provavelmente discute a qualidade dos dados como parte da avaliação das limitações dos LLMs em ambientes clínicos.
(JOTHIRAJ; MASHHADI, 2023)	50	Embora discuta privacidade e diversidade de dados, o foco está em aprendizado federado e modelos de difusão, não diretamente na qualidade dos dados para LLMs.
(VIANA et al., 2024)	0	O artigo está focado em metagenômica e não se relaciona à qualidade de dados em LLMs ou IA Generativa.

Fonte: Autor

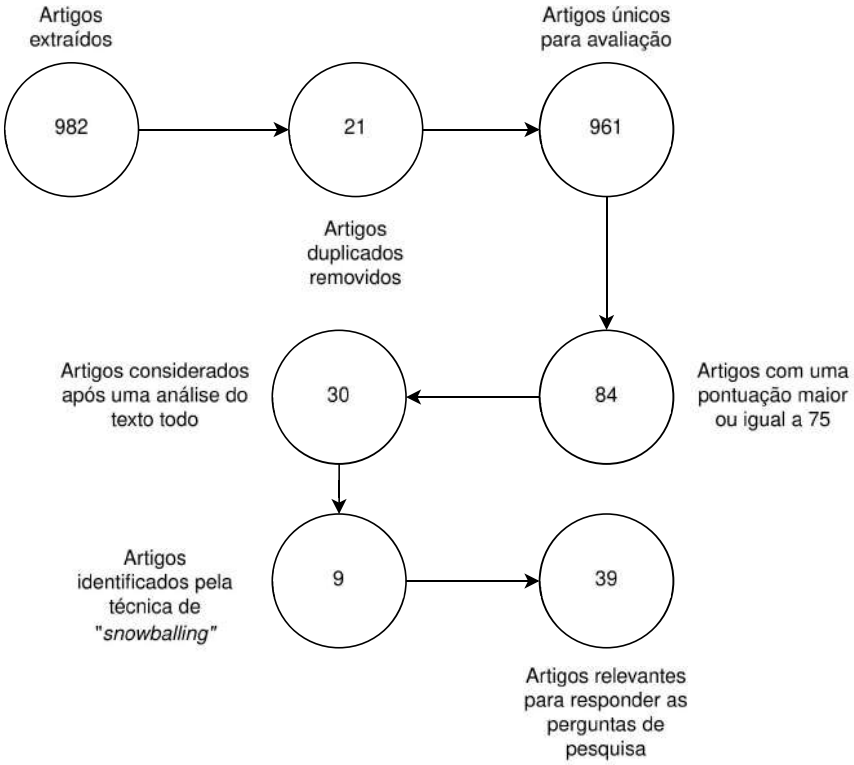
pelo modelo de linguagem eram condizentes com as expectativas. Após essa validação, foram filtrados todos os artigos com pontuação igual ou superior a 75. Tendo realizado este filtro, restaram 84 artigos para serem analisados.

Tendo reduzido o universo de artigos, foi realizada uma análise mais aprofundada, onde cada artigo foi acessado individualmente e uma leitura do conteúdo completo foi realizada, aplicando-se os critérios de exclusão. Restando assim, 30 artigos.

Também, durante essa análise aplicou-se a técnica de *snowballing* que consiste em um método para identificação de estudos adicionais relevantes por meio do exame das referências citadas em um artigo primário. Essa abordagem expande o escopo da revisão de forma iterativa, já que cada novo estudo relevante pode conter suas próprias referências que levam a outros estudos, criando um processo contínuo de descoberta e

inclusão (WOHLIN, 2014). Neste trabalho, optou-se por aplicar essa técnica apenas no primeiro nível de artigos, resultando na adição de 9 trabalhos. Resultando assim, em um universo de 39 artigos. A Figura 2 resume o processo de triagem dos artigos.

Figura 2: Funil Artigos



Fonte: Autor

3.5 Extração de Dados e Análise

O processo de extração e análise de dados foi projetado para capturar e sintetizar sistematicamente os *insights* de cada artigo selecionado, para uma melhor compreensão das práticas de qualidade de dados no desenvolvimento de LLMs. Na extração, foi utilizada uma estratégia estruturada, que orientou a recuperação consistente de informações relevantes, incluindo *frameworks*, metodologias, desafios e soluções propostas. Os dados extraídos foram subsequentemente categorizados e analisados para identificar temas abrangentes, tendências e lacunas na literatura.

3.6 Avaliação da Qualidade dos Artigos Selecionados

Tendo em vista um maior refinamento das publicações encontradas, foram definidos alguns critérios de qualidade a serem aplicados às publicações. Esses critérios são evidenciados na Tabela 6.

Tabela 6: Critérios de qualidade

	Critérios
1	Existe uma descrição adequada do contexto em que o estudo foi realizado?
2	O metodologia da pesquisa foi adequado para atender os objetivos da pesquisa?
3	A estratégia de seleção de dados foi adequada aos objetivos da pesquisa?
4	Os dados foram coletados de maneira adequada para responder as questões?
5	A análise dos dados foi suficientemente rigorosa?
6	Há uma descrição clara dos resultados?
7	O estudo possui valor para a academia ou para a indústria?

Fonte: Adaptado de (COELHO, 2022)

Para cada critério, ao analisar o artigo, é atribuída uma pontuação, de acordo com a resposta para cada critério. As pontuações e respostas possíveis estão detalhadas na Tabela 7.

Tabela 7: Peso resposta critérios de qualidade

	Resposta	Pontuação
1	O artigo atende ao critério avaliado ou o critério não se aplica.	1.0
2	O artigo não deixa claro se atende ou não ao critério.	0.5
3	Não existe nada no artigo que atenda ao critério avaliado.	0.0

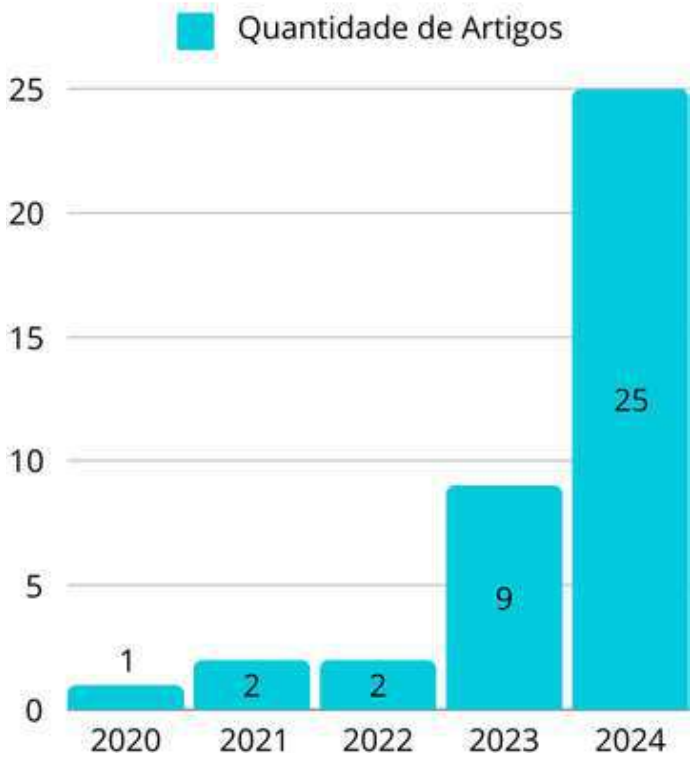
Fonte: (COELHO, 2022)

Dada a pontuação para cada artigo, conforme detalhado no Apêndice A, foram selecionados todos aqueles com uma pontuação maior que 4.5. Como a menor pontuação foi 6.0, nenhum artigo foi excluído.

4 RESULTADOS

Do universo de 961 artigos, apenas 39, ou seja 4,02% das publicações retornadas discorrem sobre qualidade de dados no desenvolvimento de LLMs. Na Figura 3 temos uma distribuição dos artigos selecionados ao longo dos anos. Apesar da data de corte definida na seção de metodologia, não foi encontrado nenhum artigo anterior a 2020, o que evidencia o quão novo é o assunto. Todos os artigos são recentes, com 64,1% desses tendo sido publicados no ano de elaboração deste trabalho. Observa-se também que, entre o ano de 2023 e 2024, houve um aumento de 170% nas publicações, passando de 9 para 25. Como o levantamento dos artigos foi realizado antes do ano de 2024 finalizar, é provável que ainda mais artigos tenham sido publicados sobre o tema.

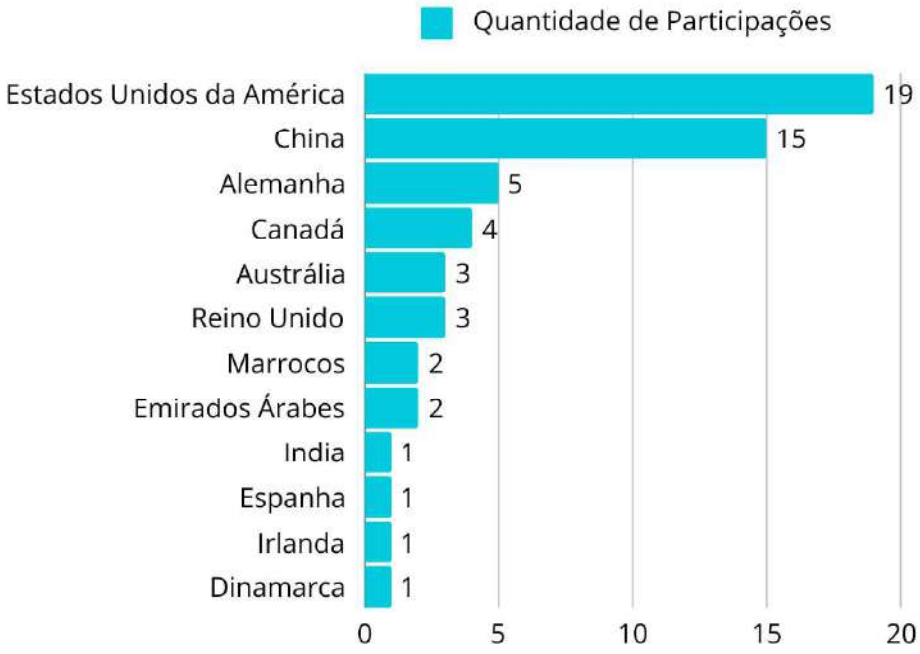
Figura 3: Publicações por Ano



Fonte: Autor

Na Figura 4, o gráfico ilustra a distribuição da participação dos países nos artigos analisados. Vale destacar que a somatória das participações excede o número total de artigos, uma vez que muitos trabalhos foram produzidos por equipes compostas por autores de diferentes nacionalidades, refletindo a natureza colaborativa e internacional da pesquisa científica.

Figura 4: Participação por País



Fonte: Autor

Observa-se uma clara predominância da participação de dois países: Estados Unidos e China. Juntos, eles são responsáveis por 59% das contribuições registradas, com 19 e 15 participações, respectivamente. Esse domínio reflete o forte investimento e a liderança dessas nações no desenvolvimento de tecnologias avançadas, incluindo LLMs. Além disso, países como Alemanha (5 participações), Canadá (4), Austrália (3) e Reino Unido (3) também desempenham papéis importantes, embora em menor escala. Esses números sugerem que, embora a pesquisa seja liderada por grandes potências científicas, há uma contribuição de outras nações, principalmente da Europa e da Oceania. Países como Marrocos e Emirados Árabes, cada um com duas participações, destacam-se como representantes do Oriente Médio e do Norte da África, indicando o interesse emergente dessas regiões no tema. Por outro lado, países como Índia, Espanha, Irlanda e Dinamarca, com uma participação cada, reforçam a ideia de que, embora a produção científica seja concentrada, há um alcance global no desenvolvimento desses estudos.

Os dados refletem tanto a concentração da pesquisa em polos tradicionais de ciência e tecnologia quanto a participação de outros países, evidenciando a crescente internacionalização e colaboração científica no campo dos LLMs.

4.1 Como a qualidade de dados impacta o desempenho e confiança de LLMs?

Qualidade de dados surgiu como um pilar para garantir um bom desempenho e confiança em LLMs. Como esses modelos têm se tornado parte integrante de várias aplicações práticas, a qualidade dos dados de treinamento é fundamental para a qualidade da capacidade de generalização em diversos contextos.

Qualidade de dados se correlaciona diretamente com o desempenho e confiabilidade dos LLMs, como evidenciado em Zhou et al. (2024b). Esse estudo destaca dimensões como consistência, completude e representatividade, que são cruciais para manter a eficácia dos modelos em diferentes tarefas. Variabilidade ou insuficiências nessas dimensões degradam o desempenho dos modelos, reforçando assim o quão influente é a qualidade de dados nos resultados dos LLMs.

Uma curadoria eficiente dos dados e fases de pré-processamento dos dados melhoram a confiabilidade dos LLMs ao endereçar problemas como ruído, desbalanceamento, e duplicações nas bases de treinamento. Tran et al. (2022) analisam o impacto de uma preparação meticulosa dos dados para garantir confiabilidade, enquanto em outro trabalho, Zhang et al. (2024) enfatizam que o pré-processamento é uma fase crítica para eliminar vieses e erros, assim, melhorando a acurácia de modelos de linguagem.

LLMs são sensíveis a variações contextuais nas bases de treinamento. Em Karra e Lasfar (2024), é demonstrado como inconsistências nos dados de entrada dos modelos podem levar a resultados imprevisíveis, principalmente em sistemas que dependem de tarefas de perguntas e respostas. Tais descobertas ressaltam a necessidade de alinhar as características dos dados com os requisitos específicos do modelo.

Embora LLMs, como o *ChatGPT*, sejam treinados em volumes massivos de dados, Li et al. (2024b) demonstraram a importância de priorizar bases de dados de alta qualidade em detrimento de grandes bases de dados generalistas para melhorar o desempenho de LLMs. No mesmo estudo, os autores observaram que a dependência excessiva de grandes volumes de dados de baixa qualidade frequentemente compromete a confiabilidade do modelo. Esse cenário sugere que, em vez de priorizar apenas a quantidade, o foco em bases

de dados validadas e direcionadas pode proporcionar uma generalização mais robusta e precisa dos modelos de linguagem.

Banco de dados de domínios específicos com alta qualidade provêm oportunidades para otimizar LLMs em áreas especializadas. Bojic et al. (2023) ilustram como domínios específicos de dados devidamente validados podem melhorar a confiabilidade dos modelos e reduzir a variabilidade dos mesmos.

A governança e gerenciamento de dados em escala, como detalhado por Wang et al. (2024), enfatiza a importância de *frameworks* e ferramentas para lidar com grandes bases de dados de forma sistemática. Gerenciamento robusto de dados garante a acessibilidade, consistência e qualidade de dados no ciclo de vida do treinamento de LLM. Além disso, o artigo realça a necessidade de *pipelines* de dados automáticos e escaláveis para um monitoramento e melhoria contínua.

O trabalho de Jernite et al. (2022) explora a interseção entre governança de dados, considerações éticas e desafios operacionais. Destaca a necessidade de práticas de dados transparentes e justas, alinhadas a preocupações mais amplas sobre privacidade de dados, mitigação de vieses e confiança pública. O artigo enfatiza princípios como controles rigorosos de acesso a dados, documentação abrangente e rastreabilidade como estratégias críticas para enfrentar esses desafios enquanto melhora a qualidade dos dados.

Os artigos, de forma coletiva, enfatizam que o desempenho e confiabilidade dos LLMs estão intrinsecamente ligados à qualidade de dados. Através dos artigos analisados, pode-se perceber alguns padrões:

- **Dependência dimensional**

  O desempenho de LLMs é um problema multi-facetado e que exige avaliações abrangentes da dimensões de qualidade. As dependências dimensionais descrevem os aspectos interconectados da qualidade dos dados — como completude, precisão, consistência, representatividade, equidade — que, em conjunto, influenciam o desempenho e a confiabilidade dos LLMs. Melhorar uma dimensão frequentemente impacta as outras, exigindo uma abordagem equilibrada e integrada para evitar erros em cascata e maximizar a robustez, equidade e sustentabilidade do modelo.

- **Redução de erros**

  Através do trabalhos analisados, evidencia-se que preprocessamento dos dados de treinamento reduzem erros e vieses, o que é crítico para atingir uma boa confiabilidade nos modelos.

- **Foco no Contexto**

  A clareza, estrutura e variabilidade das informações contextuais afetam de forma crítica o desempenho e a confiabilidade dos LLMs, especialmente em tarefas como de perguntas e respostas. Contextos de alta qualidade — livres de ambiguidades, erros ou inconsistências — melhoram a compreensão e reduzem erros de resposta em até 27%. Dados contextuais estruturados e livres de ruídos garantem que os modelos possam interpretar e responder aos *inputs* de forma confiável, destacando a necessidade de conjuntos de dados bem validados.

- **Estratégias de boas bases de dados**

  A mudança do enfoque de quantidade para qualidade, destaca uma tendência crescente em direção à eficiência e precisão na preparação de conjuntos de dados.

- **Escalabilidade e Governança**

  Necessidade de sistemas escaláveis e automáticos para gerenciamento de dados, garantindo assim a qualidade e desempenho do modelo.

- **Impactos sociais e éticos**

  Os trabalhos mostram a importância de práticas éticas de dados e estruturas de governança em mitigar riscos e melhorar a transparência.

A síntese das pesquisas ressalta que a qualidade dos dados não é apenas um fator de suporte, mas um determinante fundamental para o desempenho e a confiabilidade dos LLMs. Ao abordar as dimensões da qualidade dos dados, focar nas necessidades específicas de cada domínio e priorizar dados de alta qualidade em vez de grandes volumes, os profissionais podem melhorar os resultados dos LLMs. Além disso, a inclusão de *frameworks* robustos de governança de dados garante práticas sustentáveis e éticas no gerenciamento de grandes conjuntos de dados, atendendo às preocupações sociais junto com os requisitos técnicos.

4.2 Quais são os desafios e soluções existentes para a qualidade de dados no treinamento de LLMs?

O treinamento de LLMs depende de volumes massivos de dados, tornando a qualidade desses dados um fator crucial para a efetividade dos modelos. Dados de alta qualidade não apenas melhoram a capacidade de generalização dos modelos, mas também reduzem

vieses, aumentam a acurácia e ampliam a confiabilidade dos resultados. Entretanto, as bases de dados empregadas no treinamento dessa classe de modelos apresentam não apenas uma vasta diversidade de conteúdos e formatos, mas também um volume significativo. Esses fatores introduzem desafios consideráveis para garantir consistência, diversidade e acurácia nos dados utilizados. A heterogeneidade das fontes pode levar a inconsistências e vieses nos modelos, comprometendo a confiabilidade das aplicações.

4.2.1 Desafios em qualidade de dados para LLMs

Garantir a qualidade de dados é fundamental para o desenvolvimento de LLMs, pois influencia diretamente o desempenho, capacidades de generalização e equidade. Equidade refere-se à representação justa de grupos e perspectivas diversas nos conjuntos de dados de treinamento, garantindo resultados imparciais e confiáveis para todos os perfis de usuários. Entretanto, a quantidade e diversidade das bases de dados utilizadas no treinamento introduzem vários desafios. Desde tratar ruído e redundância até lidar com viés, a qualidade de dados continua sendo um gargalo no processo de otimizar as capacidades de LLMs. Na Figura 5 temos uma nuvem de palavras que realça os principais desafios encontrados entre os artigos analisados.

Figura 5: Nuvem Palavras Desafios Qualidade Dados Treinamento LLMs

Fonte: Autor

4.2.1.1 Ruído, Redundância e Escala

Grandes bases de dados frequentemente contêm ruído e/ou informações redundantes, o que diminui a efetividade dos dados no treinamento de LLM. O trabalho de Longpre et

al. (2023) relata como dados duplicados e com pouca qualidade afetam de forma negativa a eficiência do treinamento, aumentam as chances de falsas correlações e padrões enganosos no resultado dos modelos. Ainda, a propagação de ruído durante o treinamento aumenta o risco de sobreajuste das características irrelevantes, que são elementos no conjunto de dados que não contribuem significativamente para o desempenho preditivo do modelo ou para os objetivos específicos da tarefa. A escala absoluta de bases de dados modernas amplificam esses problemas, como é discutido em Chen et al. (2024), onde são apresentados os desafios computacionais e logísticos para táticas de pré-processamento, filtragem e deduplicação em grandes bases de dados. Apesar de utilizar sistemas de pré-processamento modulares e escaláveis, operações que demandam muitos recursos continuam sendo um obstáculo para garantir a uniformidade e a limpeza dos dados de treinamento.

Para somar à complexidade do problema, em Shankar et al. (2024), os autores destacam a necessidade de sintetizar asserções de qualidade para os dados a fim de minimizar a redundância de instruções e garantir o alinhamento com os critérios de tarefas específicas. A integração de tais processos é crucial para redução de ruído enquanto mantêm a integridade semântica das bases de dados de treinamento.

4.2.1.2 Viés e falta de representação

Viés em bases de dados, principalmente aquelas derivadas de fontes extraídas da internet, pode propagar nos LLMs e produzirem resultados injustos e não confiáveis. Em Zhou et al. (2024a), os autores discutem a prevalência de vieses culturais e demográficos, o que compromete a equidade e a inclusividade das aplicações resultantes. Esse desafio é amplificado pela falta de representação diversificada nos dados de treinamento, levando a distorções sistêmicas nas previsões do modelo. Em Yu et al. (2024), os autores enfatizam que uma representação inadequada de diferentes perspectivas e contextos limita a generalização de LLMs. Além disso, vieses frequentemente emergem por casa de alguma super-representação de domínios específicos ou demográficos, perpetuando assim tratamento desbalanceado entre grupos.

O trabalho de Li et al. (2024a), endereça esses desafios ao aplicar métricas IFD para identificar e reduzir o ruido em bases de dados de instruções. Essa abordagem ajuda a melhorar a compatibilidade entre os dados e as necessidades da tarefa solicitada enquanto mitiga desequilíbrios representativos.

4.2.1.3 Problemas com dados de domínios específicos e instruções

Bases de dados de instruções, utilizadas para treinar modelos em contexto específicos, geralmente enfrentam desafios de ambiguidade, inconsistência, desalinhamento, o que impede as capacidades de seguir instruções dos LLMs. Liu et al. (2024) argumentam que pares de instrução-resposta sem uma curadoria adequada prejudicam o alinhamento do modelo e contribuem para erros como alucinações. Bases de dados de domínios específicos, incluindo aqueles como de medicina e/ou programação, apresentam uma complexidade adicional. No trabalho de Rozière et al. (2024), os autores revelam como dados de código com ruído e não estruturados afetam o desempenho dos modelos em aplicações especificas. Em Zhang et al. (2024) são identificados problemas como formatos caóticos, problemas de privacidade e textos de baixa qualidade em dados médicos, o que requer soluções de pré-processamento especificas. Abordar esses desafios exigem *frameworks* próprios de domínio que combinem técnicas de preservação da privacidade com estratégias avançadas de redução de ruído.

4.2.1.4 Escalabilidade da curadoria e filtros

O crescimento exponencial das bases de dados utilizadas no treinamento de LLMs apresentam desafios de escalabilidade para garantir a qualidade. Penedo et al. (2023) demonstram que mesmo com esforços significativos e regidos para deduplicação, bases de dados na escala da internet são propensos a inconsistências e, portanto, requerem uma curadoria extensiva para manter a coerência e relevância.

O *framework IterClean*, proposta por Ni et al. (2024), aborda o problema de escalabilidade ao aplicar métodos iterativos de limpeza de dados que usam modelos de *feedback* para identificar e ratificar ruído ou dados redundantes. Essa abordagem é particularmente valiosa para bases de dados com um número limitado de dados com marcações, onde a curadoria manual é impraticável. De forma similar, a arquitetura modular *Data-Juicer*, proposta por Chen et al. (2024), propõe uma infraestrutura para processamento e deduplicação de forma escalável, mas a dependência da solução em processos de computação intensiva evidencia a necessidade de soluções que também sejam eficientes no uso de recursos. Apesar desses avanços, manter o equilíbrio entre escalabilidade e qualidade continua sendo um obstáculo importante. Como as bases de dados continuam a crescer em tamanho e complexidade, *frameworks* devem evoluir para incorporar soluções automatizáveis e adaptativas sem comprometer acurácia e diversidade.

4.2.2 Soluções para melhorar a qualidade de dados

Endereçar desafios na qualidade de dados para LLMs requer abordagens inovadoras e sistemáticas construídas especificamente para as demandas únicas de grandes bases de dados e de domínio específico. A literatura analisada apresenta uma série de *frameworks* e ferramentas projetadas para mitigar problemas como ruído, redundância, viés, e escalabilidade enquanto melhora a diversidade das bases de dados e o alinhamento com os objetivos do modelo. Essas soluções utilizam tanto processos automatizados quanto expertise de domínio para otimizar a preparação de dados, garantindo maior precisão e robustez nos resultados dos LLMs. Na Tabela 8 temos um resumo das principais ferramentas proposta na literatura, listando os principais atributos e contexto de aplicação.

Tabela 8: Soluções e *Frameworks* para Qualidade de Dados em LLMs

Solução/Framework	**Principais Atributos**	**Contexto de Aplicação**
EasyInstruct Framework (OU et al., 2024)	Estrutura modular para geração de instruções, métricas avançadas (GPTScore, perplexidade), geração automatizada de instruções.	Garante conjuntos de instruções de alta qualidade e diversidade; reduz redundância; melhora o alinhamento do modelo.
CoachLM (LIU et al., 2024)	Revisão automática de instruções de baixa qualidade, treinamento orientado por especialistas, avaliação de qualidade em nove dimensões.	Melhora as capacidades de seguir instruções; mitiga erros; garante diversidade no conjunto de dados.
Data-Juicer (CHEN et al., 2024)	Mais de 50 operadores modulares de pré-processamento; suporte para processamento de dados escalável e flexível.	Ideal para pré-processamento em larga escala; garante dados de treinamento livres de ruído; permite experimentação com receitas de dados.

Continua na próxima página

Tabela 8 – continuação da página anterior

Solução/Framework	Principais Atributos	Contexto de Aplicação
Oasis System (ZHOU et al., 2024a)	Deduplicação adaptativa, filtragem modular baseada em regras, filtros neurais sem viés, métricas de avaliação holísticas.	Melhora a fluência, coerência e diversidade do corpus; reduz vieses em conjuntos de dados de grande escala.
IterClean (NI et al., 2024)	Limpeza iterativa usando feedback do modelo; framework modular para lidar com erros.	Melhora incrementalmente a qualidade dos dados; aumenta a robustez; útil para conjuntos de dados rotulados limitados.
SPADE (SHANKAR et al., 2024)	Sintetiza afirmações de qualidade de dados; reduz afirmações redundantes; utiliza deltas de prompts para refinamento.	Garante conformidade com critérios de qualidade; reduz alucinações e erros de instruções.
Selective Reflection-Tuning (LI et al., 2024a)	Colaboração professor-aluno; métricas de Dificuldade de Seguir Instruções (IFD) para refinamento direcionado.	Melhora a compatibilidade instrução-resposta; reduz ruído; otimiza a relevância dos dados.
Comprehensive Medical Framework (ZHANG et al., 2024)	Pré-processamento em quatro módulos: unificação de formato, filtragem de qualidade, deduplicação, redução de privacidade.	Garante dados médicos compatíveis com privacidade e livres de ruído; melhora os resultados de treinamento específicos ao domínio.

Continua na próxima página

Tabela 8 – continuação da página anterior

Solução/Framework	Principais Atributos	Contexto de Aplicação
LLM-Assisted Code Cleaning (JAIN et al., 2024)	Modularização de código, renomeação de variáveis, verificação de equivalência funcional.	Otimiza conjuntos de dados para LLMs de geração de código; reduz vieses e erros de saída.
RefinedWeb Dataset (PENEDO et al., 2023)	Pipeline de Refinamento MacroData; filtragem adaptativa para diversidade e coerência de conteúdo.	Prepara dados web de alta qualidade; garante melhorias de desempenho em zero-shot.

4.2.2.1 Síntese

A análise dos desafios e soluções revelam alguns padrões e percepções:

- **Abordagens modulares e iterativas**

  Várias ferramentas, incluindo *IterClean* e *Data-Juicer*, enfatizam uma abordagem iterativa e modular para o refinamento, o que abre portas para processamento adaptativo e escalar dos dados de treinamento.

- ***Trade-Offs* entre desempenho e mitigação de viés**

  Técnicas de filtragem para qualidade e viés frequentemente envolvem alguns *trade-offs*, como relatado por Longpre et al. (2023), onde reduzir toxidade nos dados geralmente compromete as capacidade de generalização dos modelos.

- **Soluções demandam recursos**

  O volume das bases de dados utilizadas no treinamento de LLMs necessitam de uma quantidade significativa de recursos computacionais e humanos para filtragem, deduplicação e adaptação para domínios específicos, como é demonstrado por Penedo et al. (2023).

- **Desafios de domínios específicos**

Soluções construídas para tarefas específicas, tais como medicina ou programação, realçam a importância de pre-processamento que levem em conta o contexto do dados e *fine-tuning*.

Apesar das soluções apresentadas endereçarem diversos dos desafios, várias lacunas permanecem, tais como métricas padronizadas de qualidade de dados e desenvolvimento de aplicações que podem ser utilizadas em diversos cenários. As soluções que envolvem a integração da expertise humana com processos automatizados são caminhos promissores para resolver essas lacunas.

Por fim, garantir a qualidade dos dados é fundamental para o treinamento de LLM, influenciando na confiabilidade, justiça e versatilidade. Desafios como ruído, viés e escalabilidade demandam soluções inovadoras incluindo sistemas de preprocessamento modular, processos de limpeza iterativa e *frameworks* para domínios específicos. Apesar deste trabalho ter levantado abordagens promissoras, mais pesquisas são necessárias para definir métricas padronizadas de qualidade e o equilíbrio entre qualidade, mitigação de viés e escalabilidade.

4.3 Como os problemas de qualidade de dados contribuem para vieses e erros em inteligências artificiais generativas?

Dados com pouca qualidade podem introduzir vieses e até amplificar os existentes, propagar erros e assim comprometer a confiabilidade, impactando assim as aplicações éticas e sociais de IA.

Erros de anotação e dados rotulados incorretamente inserem, diretamente, vieses em sistemas de IA. Por exemplo, quando as bases de dados utilizadas em treinamento se baseiam em anotações inconsistentes e/ou incompletas, LLMs falham em generalizar, e frequentemente replicam erros com confidência, isto é, passam a ideia de informação correta, porém, quando analisado é falsa. Um exemplo apresentado por Tran et al. (2022) mostra que dados rotulados incorretamente em tarefas de detecção de intrusão fez com que os modelos identificassem, erroneamente, atividades benignas como maliciosas, evidenciando assim o efeito cumulativo desses vieses na tomada de decisões críticas. Detecção de intrusão, no contexto do artigo, refere-se ao processo de monitoramento e análise do tráfego de rede ou das atividades do sistema para identificar acessos não autorizados, violações ou atividades maliciosas.

Vieses são perpetuados quando bases de dados não representam, de forma proporcional, todos os grupos presentes no contexto da aplicação. Trabalhos como o de Zhou et al. (2024a) descrevem como a sub-representação de grupos minoritários da sociedade e, portanto, geram resultados enviesados ou com esteriótipos. Zhang et al. (2024) descrevem como bases de dados recuperadas da internet geralmente representam, de forma desproporcional, doenças mais comuns em centros urbanos e sub-representam problemas de saúde na área rural. Esse desbalanceamento nos dados de treinamento afeta a habilidade do LLM de gerar análises médicas igualmente acuradas.

Ruído nos dados, tais como registros duplicados, e dados que não são relevantes, ou ainda, formatos incorretos, introduzem imprecisões no resultado de IAs. Shankar et al. (2024) detalham que erros como entradas repetidas resultam em sobreajuste, fazendo com que os LLMs priorizem padrões repetitivos ao invés de novas generalizações para um dado contexto. Em áreas específicas, como é o caso de dados médicos, bases de dados com ruído e com terminologia inconsistente fazem com que os modelos de linguagem deem conselhos contraditórios baseados em variações mínimas na forma de como a pergunta é formulada.

As bases de dados que não têm uma variabilidade linguística muito boa podem gerar modelos que têm seu desempenho minado quando configurados para realizar tarefas em várias línguas ou até falham em entender dialetos que são sub-representados (ou não presentes) na base de dados. Karra e Lasfar (2024) discutem como dados de treinamento contextualmente limitados levam sistemas de pergunta e resposta a gerar resultados menos acurados em casos de cenários complexos de linguagem, revelando vieses contra estruturas sintáticas mais intrincadas.

Vieses já existentes na sociedade estão incorporados nas bases de treinamento e são amplificados, e até perpetuados por LLMs, como apresentado no trabalho de Li et al. (2024a), no qual é demonstrado que *prompts* enviesados em bases de dados de instruções ensinaram o modelo a favorecer certos grupos demográficos em contextos como candidaturas a empregos ou avaliações acadêmicas. De forma similar, em Garrido-Muñoz, Martínez-Santiago e Montejo-Ráez (2023), os autores evidenciam como LLMs treinados em bases de dados em espanhol apresentam um viés sistemático herdado dos dados de treinamento. Por exemplo:

- Modelos masculinos, como "Ele é o mais [MASK]", produzem descritores como "inteligente"e "forte", alinhando-se a estereótipos de liderança e intelectuais.
- Modelos femininos, por outro lado, geram adjetivos como "bonita"e "emocional", perpetuando tendências sociais de focar na aparência e nos atributos emocionais das

mulheres em detrimento de suas capacidades

A reprodução constante de vieses sociais aumenta as desigualdades em aplicações como ferramentas de recrutamento, onde homens podem ser preferidos de forma injusta para cargos de liderança, ou em contextos educacionais, onde as conquistas de mulheres podem receber menos valor. Assim como acontece com os conjuntos de dados usados para instruções, os vieses incorporados nesses modelos de linguagem criam ciclos que reforçam estereótipos em diferentes tarefas, mostrando a necessidade urgente de reduzir esses vieses durante a curadoria dos dados.

Dados de treinamento ambíguos inserem incertezas em inteligências artificiais generativas. Liu et al. (2024) mostram que modelos treinados em bases de dados com instruções incompletas frequentemente geram respostas vagas ou enganosas quando confrontados com tarefas que exigiam direções específicas.

Em tarefas como geração de código, erros nos dados de treinamento leva a falhas encadeadas. Em Jain et al. (2024) é demonstrado que dados com códigos de baixa qualidade resultam em LLMs produzindo código que são sintaticamente corretos, porém com funcionalidades incorretas, o que poderia propagar para sistemas em produção levando a sérias consequências.

Essas evidencias enfatizam que problemas de qualidade de dados são fundamentais para a geração e propagação de vieses e erros em inteligências artificiais generativas. Baseados nos artigos, foram identificados alguns padrões no tema:

- **Proliferação de viés por problemas sistêmicos nos dados**:

  Dados de baixa qualidade são um fator significativo na reprodução e amplificação de estereótipos sociais. Por exemplo, Shankar et al. (2024) destacam como conjuntos de dados enviesados por gênero e cultura levam os LLMs a perpetuar estereótipos prejudiciais, como associar determinadas profissões a gêneros específicos ou normas culturais. Da mesma forma, conjuntos de dados médicos com vieses centrados em áreas urbanas distorcem as necessidades de saúde rural, resultando em saídas de modelos inequitativas, como observado em Zhang et al. (2024).

- **Ciclos de retroalimentação em modelos de IA generativa**:

  Modelos de IA generativa criam, ciclos de retroalimentação que reforçam os próprios vieses quando treinados com dados enviesados ou incompletos. Por exemplo, o trabalho de Li et al. (2024a) demonstra como os vieses introduzidos por conjuntos de

dados de instruções propagam erros em iterações subsequentes do modelo, agravando o viés em vez de mitiga-lo. Esse ciclo compromete a equidade e a adaptabilidade dos sistemas de IA no longo prazo.

- **Impacto de dados duplicados e ruídos**:

  Redundâncias e ruídos aumentam os erros nas saídas dos modelos. Como descrito por Shankar et al. (2024), padrões repetitivos são priorizados em detrimento de generalizações diversas devido a dados sobrepostos. Da mesma forma, Jain et al. (2024) destacam os efeitos em cascata de dados de código com ruídos, resultando em resultados funcionalmente errados em aplicações críticas.

- **Viés contextual e linguístico**:

  Lacunas linguísticas e contextuais nos conjuntos de dados criam assimetrias no desempenho dos modelos, especialmente em ambientes multilíngues ou complexos. Trabalhos como o de Karra e Lasfar (2024) enfatizam que dados de treinamento com falta de contexto levam a erros desproporcionais em cenários sintáticos mais complexos.

- **Ambiguidade e erros de anotação**:

  Dados ambíguos ou conjuntos de dados mal anotados afetam significativamente a confiabilidade dos modelos. Liu et al. (2024) mostram que instruções incompletas ou inconsistentes resultam em saídas vagas e propensas a erros. Esses problemas agravam as imprecisões dos modelos e minam a confiança nos sistemas generativos.

Conclui-se que abordar os vieses de maneira holística requer a priorização da diversidade de dados, o alinhamento contextual e a garantia da qualidade das anotações em todas as etapas da pipeline de desenvolvimento. Nesse contexto, destaca-se a importância de pesquisas futuras voltadas para o aprimoramento de ferramentas automatizadas de avaliação da qualidade dos dados e para a implementação de protocolos robustos de validação, fundamentais para mitigar vieses em todas as fases de treinamento e aplicação dos modelos.

5 CONCLUSÃO

A revisão de escopo da literatura realizada destacou que a qualidade dos dados é essencial para o desenvolvimento e a confiabilidade de grandes modelos de linguagem. Dados de baixa qualidade podem introduzir vieses e erros significativos, comprometendo a utilidade e a ética dos sistemas baseados em IA. Estratégias como a curadoria de dados, o uso de *frameworks* escaláveis e a governança robusta foram identificadas como abordagens eficazes para lidar com esses desafios.

Além disso, a pesquisa mostrou que um equilíbrio entre quantidade e qualidade dos dados pode ser alcançado para otimizar o desempenho dos modelos. A adoção de técnicas avançadas, como limpeza interativa de dados e o uso de ferramentas de revisão automatizada, também se mostrou promissora. Contudo, há uma necessidade contínua de pesquisas voltadas para a padronização e a aplicação ética no uso de dados para LLM.

Pesquisas futuras devem se concentrar no desenvolvimento de métricas padronizadas de qualidade, permitindo que os profissionais que trabalham com LLMs analisem seus dados de maneira reproduzível e monitorem a qualidade dos dados em aplicações implementadas. Existe uma necessidade crítica de ferramentas robustas para a detecção de vieses, que combinem métodos baseados em IA com supervisão humana para identificar e mitigar efetivamente os vieses nos conjuntos de dados. Além disso, as ferramentas de agentes de IA utilizadas para aumento de dados e melhoria de qualidade devem incorporar métodos de curadoria e documentação transparente, garantindo que não introduzam dados ruins ou enviesados nos conjuntos de dados que estão corrigindo.

REFERÊNCIAS

ALBALAK, A. et al. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024. Disponível em: ⟨https://arxiv.org/abs/2402.16827⟩.

BARIKERI, S. et al. Redditbias: A real-world resource for bias evaluation and debiasing of conversational language models. *arXiv preprint arXiv:2106.03521*, 2021. Disponível em: ⟨https://arxiv.org/abs/2106.03521⟩.

BATINI, C.; LENZERINI, M.; NAVATHE, S. B. A comparative analysis of methodologies for database schema integration. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 18, n. 4, p. 323–364, dez. 1986. ISSN 0360-0300. Disponível em: ⟨https://doi.org/10.1145/27633.27634⟩.

BECK, J. Quality aspects of annotated data: A research synthesis. *AStA Wirtschafts- und Sozialstatistisches Archiv*, Springer Science and Business Media LLC, v. 17, p. 331–353, 11 2023. ISSN 1863-8163. Disponível em: ⟨http://dx.doi.org/10.1007/s11943-023-00332-y⟩.

BLEIHOLDER, J.; NAUMANN, F. Data fusion. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 41, n. 1, jan. 2009. ISSN 0360-0300. Disponível em: ⟨https://doi.org/10.1145/1456650.1456651⟩.

BOJIC, I. et al. A data-centric framework for improving domain-specific machine reading comprehension datasets. In: S., T. et al. (Ed.). *ACL 2023 - 4th Workshop on Insights from Negative Results in NLP, Proceedings*. Association for Computational Linguistics (ACL), 2023. p. 19 – 32. ISBN 978-195942949-4. Cited by: 1. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174874954&partnerID=40&md5=000e79981ba5c3193a6417d2dca0f947⟩.

BROWN, T. B. et al. Language models are few-shot learners. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2020. (NIPS '20). ISBN 9781713829546.

BRUCKNER, R. M.; LIST, B.; SCHIEFER, J. Striving towards near real-time data integration for data warehouses. In: KAMBAYASHI, Y.; WINIWARTER, W.; ARIKAWA, M. (Ed.). *Data Warehousing and Knowledge Discovery*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. p. 317–326. ISBN 978-3-540-46145-6.

CHEN; CHIANG; STOREY. Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, v. 36, p. 1165, 2012. ISSN 02767783.

CHEN, D. et al. Data-juicer: A one-stop data processing system for large language models. In: *Companion of the 2024 International Conference on Management of Data*. Association for Computing Machinery, 2024. p. 120–134. ISBN 9798400704222. Disponível em: ⟨https://doi.org/10.1145/3626246.3653385⟩.

CODD, E. F. A relational model of data for large shared data banks. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 13, n. 6, p. 377–387, jun. 1970. ISSN 0001-0782. Disponível em: ⟨https://doi.org/10.1145/362384.362685⟩.

COELHO, J. *Uso de mineração de processos no mercado financeiro: uma revisão sistemática da literatura*. Monografia (Trabalho de conclusão de curso (Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas)) — Faculdade de Tecnologia de São Paulo, 2022. Disponível em: ⟨http://ric.cps.sp.gov.br/handle/123456789/10526⟩.

CôTé, P.-O. et al. Data cleaning and machine learning: a systematic literature review. *Automated Software Engineering*, Springer Science and Business Media LLC, v. 31, 6 2024. ISSN 1573-7535. Disponível em: ⟨http://dx.doi.org/10.1007/s10515-024-00453-w⟩.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: BURSTEIN, J.; DORAN, C.; SOLORIO, T. (Ed.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: ⟨https://aclanthology.org/N19-1423⟩.

DU, F. et al. A survey of llm datasets: From autoregressive model to ai chatbot. *Journal of Computer Science and Technology*, Springer Science and Business Media LLC, v. 39, p. 542–566, 5 2024. ISSN 1860-4749. Disponível em: ⟨http://dx.doi.org/10.1007/s11390-024-3767-3⟩.

ENGLISH, L. P. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. USA: John Wiley & Sons, Inc., 1999. ISBN 0471253839.

FILIPPOVA, K. Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873*, 2020. Disponível em: ⟨https://arxiv.org/abs/2010.05873⟩.

GARRIDO-MUñOZ, I.; MARTíNEZ-SANTIAGO, F.; MONTEJO-RáEZ, A. Maria and beto are sexist: evaluating gender bias in large language models for spanish. *Language Resources and Evaluation*, Springer Science and Business Media LLC, 7 2023. ISSN 1574-0218. Disponível em: ⟨http://dx.doi.org/10.1007/s10579-023-09670-3⟩.

GUNASEKAR, S. et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023. Disponível em: ⟨https://arxiv.org/abs/2306.11644⟩.

HAGER, P. et al. Evaluating and mitigating limitations of large language models in clinical decision making. *medRxiv*, Cold Spring Harbor Laboratory Press, 2024. Disponível em: ⟨https://www.medrxiv.org/content/early/2024/01/26/2024.01.26.24301810⟩.

JAIN, N. et al. Llm-assisted code cleaning for training accurate code generators. In: *12th International Conference on Learning Representations, ICLR 2024*. International Conference on Learning Representations, ICLR, 2024. Cited by: 1. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85200465118&partnerID=40&md5=714fbb529e9e4b3ba16773e84215c2c6⟩.

JERNITE, Y. et al. Data governance in the age of large-scale data-driven language technology. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery, 2022. (FAccT '22), p. 2206–2222. ISBN 9781450393522. Disponível em: ⟨https://doi.org/10.1145/3531146.3534637⟩.

JOTHIRAJ, F. V. S.; MASHHADI, A. Phoenix: A federated generative diffusion model. *arXiv preprint arXiv:2306.04098*, 2023. Disponível em: ⟨https://arxiv.org/abs/2306.04098⟩.

KANG, H. J. et al. Human-in-the-loop synthetic text data inspection with provenance tracking. In: K., D.; H., G.; S., B. (Ed.). *Findings of the Association for Computational Linguistics: NAACL 2024 - Findings*. Association for Computational Linguistics (ACL), 2024. p. 3118 – 3129. ISBN 979-889176119-3. Cited by: 0. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85197893574&partnerID=40&md5=718a39878f999486ed8fd9b6011b694d⟩.

KAPLAN, J. et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. Disponível em: ⟨https://doi.org/10.48550/arXiv.2001.08361⟩.

KARRA, R.; LASFAR, A. Impact of data quality on question answering system performances. *Intelligent Automation and Soft Computing*, Tech Science Press, v. 35, p. 335 – 349, 2023. ISSN 10798587. Cited by: 4; All Open Access, Hybrid Gold Open Access. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85135081060&doi=10.32604%2fiasc.2023.026695&partnerID=40&md5=8e65680d6b4edb56a9d3ad09514c48e1⟩.

KARRA, R.; LASFAR, A. Analysis of qa system behavior against context and question changes. *International Arab Journal of Information Technology*, Zarka Private University, v. 21, p. 191 – 200, 2024. ISSN 16833198. Cited by: 0; All Open Access, Gold Open Access. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85187480835&doi=10.34028%2fiajit%2f21%2f2%2f2&partnerID=40&md5=b66c2ccfa014eeb49c374cd9df38261e⟩.

KITCHENHAM, B. *Kitchenham, B.: Guidelines for performing Systematic Literature Reviews in software engineering. EBSE Technical Report EBSE-2007-01*. [S.l.: s.n.], 2007.

KOFOD-PETERSEN, A. How to do a structured literature review in computer science. 05 2015.

KUMAR, P. Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, Springer Science and Business Media LLC, v. 57, 8 2024. ISSN 1573-7462. Disponível em: ⟨http://dx.doi.org/10.1007/s10462-024-10888-y⟩.

LAKRETZ, Y. et al. Can transformers process recursive nested constructions, like humans? In: CALZOLARI, N. et al. (Ed.). *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022. p. 3226–3232. Disponível em: ⟨https://aclanthology.org/2022.coling-1.285⟩.

LI, M. et al. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. In: L.-W., K.; A., M.; V., S. (Ed.). *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), 2024. p. 16189 – 16211. ISBN 979-889176099-8. ISSN 0736587X. Cited by: 0. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205322984&partnerID=40&md5=5454d0704804f60e7b29ebd39eadd5f6⟩.

LI, M. et al. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In: K., D.; H., G.; S., B. (Ed.). *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*. Association for Computational Linguistics (ACL), 2024. v. 1, p. 7595 – 7628. ISBN 979-889176114-8. Cited by: 0. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85200046149&partnerID=40&md5=ce017df8ceb6d04a4bfe335ff3ee3002⟩.

LIN, Z. et al. Towards trustworthy llms: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, Springer Science and Business Media LLC, v. 57, 8 2024. ISSN 1573-7462. Disponível em: ⟨http://dx.doi.org/10.1007/s10462-024-10896-y⟩.

LIU, Y. et al. Coachlm: Automatic instruction revisions improve the data quality in llm instruction tuning. In: *Proceedings - International Conference on Data Engineering*. IEEE Computer Society, 2024. p. 5184 – 5197. ISBN 979-835031715-2. ISSN 10844627. Cited by: 0; All Open Access, Green Open Access. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85200449815&doi=10.1109%2fICDE60146.2024.00390&partnerID=40&md5=fe8c0b253bde956b4a7f395fe94e8e92⟩.

LONGPRE, S. et al. *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity*. 2023. Disponível em: ⟨https://arxiv.org/abs/2305.13169⟩.

Mendeley. *Mendeley Reference Manager*. 2024. Accessed: 2024-11-12. Disponível em: ⟨https://www.mendeley.com⟩.

MIKOLOV, T. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, v. 3781, 2013. Disponível em: ⟨https://doi.org/10.48550/arXiv.1301.3781⟩.

MYERS, D. et al. Foundation and large language models: fundamentals, challenges, opportunities, and social impacts. *Cluster Computing*, Springer Science and Business Media LLC, v. 27, p. 1–26, 11 2023. ISSN 1573-7543. Disponível em: ⟨http://dx.doi.org/10.1007/s10586-023-04203-7⟩.

NAIMI, A. I.; WESTREICH, D. J. Big data: A revolution that will transform how we live, work, and think. *American Journal of Epidemiology*, v. 179, p. 1143–1144, 5 2014. ISSN 0002-9262.

NI, W. et al. Iterclean: An iterative data cleaning framework with large language models. In: *Proceedings of the ACM Turing Award Celebration Conference - China 2024*. Association for Computing Machinery, 2024. p. 100–105. ISBN 9798400710117. Disponível em: ⟨https://doi.org/10.1145/3674399.3674436⟩.

OU, Y. et al. Easyinstruct: An easy-to-use instruction processing framework for large language models. In: CAO, Y.; FENG, Y.; XIONG, D. (Ed.). *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Bangkok, Thailand: Association for Computational Linguistics, 2024. p. 94–106. Disponível em: ⟨https://aclanthology.org/2024.acl-demos.10⟩.

PATASHNIK, Oren. *BibTeXing*. 1988. Documentation for the BibTeX program distributed with LaTeX. Disponível em: ⟨http://mirrors.ctan.org/biblio/bibtex/base/btxdoc.pdf⟩.

PENEDO, G. et al. *The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only*. 2023. Disponível em: ⟨https://arxiv.org/abs/2306.01116⟩.

PETERS, M. E. et al. Deep contextualized word representations. *ArXiv*, abs/1802.05365, 2018. Disponível em: ⟨https://api.semanticscholar.org/CorpusID:3626819⟩.

QIAN, C.; REIF, E.; KAHNG, M. Understanding the dataset practitioners behind large language models. In: *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 2024. ISBN 979-840070331-7. Cited by: 2; All Open Access, Bronze Open Access. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85194186677&doi=10.1145%2f3613905.3651007&partnerID=40&md5=f5649963f26828e81114366522a6a248⟩.

RADFORD, A. et al. *Language models are unsupervised multitask learners*. 2019. Disponível em: ⟨https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe⟩.

REDMAN, T. C. The impact of poor data quality on the typical enterprise. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 41, n. 2, p. 79–82, fev. 1998. ISSN 0001-0782. Disponível em: ⟨https://doi.org/10.1145/269012.269025⟩.

ROZIèRE, B. et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2024. Disponível em: ⟨https://arxiv.org/abs/2308.12950⟩.

SCHWAB, K. *The Fourth Industrial Revolution*. USA: Crown Publishing Group, 2017. ISBN 1524758868.

SCHWABE, D. et al. The metric-framework for assessing data quality for trustworthy ai in medicine: a systematic review. *npj Digital Medicine*, Springer Science and Business Media LLC, v. 7, 8 2024. ISSN 2398-6352. Disponível em: ⟨http://dx.doi.org/10.1038/s41746-024-01196-4⟩.

SHANKAR, S. et al. Spade: Synthesizing data quality assertions for large language model pipelines. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 17, p. 4173 – 4186, 2024. ISSN 21508097. Cited by: 0. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205301650&doi=10.14778%2f3685800.3685835&partnerID=40&md5=b5a9ed8bf5a811d33586c28f91b3ed94⟩.

SHEN, L. et al. Identifying untrustworthy samples: Data filtering for open-domain dialogues with bayesian optimization. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Association for Computing Machinery, 2021. p. 1598–1608. ISBN 9781450384469. Disponível em: ⟨https://doi.org/10.1145/3459637.3482352⟩.

SUN, Y. et al. An integrated data processing framework for pretraining foundation models. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. [s.n.], 2024. p. 2713–2718. Disponível em: ⟨https://doi.org/10.48550/arXiv.2402.16358⟩.

TOUVRON, H. et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. Disponível em: ⟨https://arxiv.org/abs/2307.09288⟩.

TRAN, N. et al. Data curation and quality evaluation for machine learning-based cyber intrusion detection. *IEEE Access*, v. 10, p. 121900–121923, 10 2022. ISSN 2169-3536.

VADAPALLI, J. et al. Incorporating citizen-generated data into large language models. In: *Proceedings of the 25th Annual International Conference on Digital Government Research*. New York, NY, USA: Association for Computing Machinery, 2024. (dg.o '24), p. 1023–1025. ISBN 9798400709883. Disponível em: ⟨https://doi.org/10.1145/3657054.3659119⟩.

VASWANI, A. et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. ISBN 9781510860964.

VIANA, P. A. B. et al. Design and implementation of a metagenomic analytical pipeline for respiratory pathogen detection. *BMC Research Notes*, v. 17, n. 1, p. 291, out. 2024.

WAND, Y.; WANG, R. Y. Anchoring data quality dimensions in ontological foundations. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 39, n. 11, p. 86–95, nov. 1996. ISSN 0001-0782. Disponível em: ⟨https://doi.org/10.1145/240455.240479⟩.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, Taylor & Francis, Ltd., v. 12, n. 4, p. 5–33, 1996. ISSN 07421222. Disponível em: ⟨http://www.jstor.org/stable/40398176⟩.

WANG, Z. et al. Data management for training large language models: A survey. *CoRR*, 2024. Disponível em: ⟨https://arxiv.org/abs/2312.01700⟩.

WOHLIN, C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, p. 1–10, 2014. Accessed: 2024-11-12. Disponível em: ⟨https://doi.org/10.1145/2601248.2601268⟩.

YU, X. et al. What makes a high-quality training dataset for large language models: A practitioners' perspective. In: *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*. Association for Computing Machinery, 2024. p. 656–668. ISBN 9798400712487. Disponível em: ⟨https://doi.org/10.1145/3691620.3695061⟩.

ZHANG, C. et al. A comprehensive data preprocessing framework towards improving internet chinese medical data quality. In: *2024 5th International Conference on Computer Engineering and Application, ICCEA 2024*. Institute of Electrical and Electronics Engineers Inc., 2024. p. 514 – 520. ISBN 979-835038677-6. Cited by: 0. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85201158461&doi=10.1109%2fICCEA62105.2024.10603802&partnerID=40&md5=bcb921e94beaaa35b7218196c0fc9f66⟩.

ZHANG, X.; ABDUL-MAGEED, M.; LAKSHMANAN, L. V. S. *Autoregressive + Chain of Thought = Recurrent: Recurrence's Role in Language Models' Computability and a Revisit of Recurrent Transformer*. 2024. Disponível em: ⟨https://arxiv.org/abs/2409.09239⟩.

ZHOU, T. et al. Oasis: Data curation and assessment system for pretraining of large language models. In: K., L. (Ed.). *IJCAI International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence, 2024. p. 8855 – 8859. ISBN 978-195679204-1. ISSN 10450823. Cited by: 0. Disponível em: ⟨https://www.scopus.com/inward/record.uri?eid=2-s2.0-85204310040&partnerID=40&md5=510e0b68724a9fe55502697369dac987⟩.

ZHOU, Y. et al. A survey on data quality dimensions and tools for machine learning invited paper. In: *2024 IEEE International Conference on Artificial Intelligence Testing (AITest)*. [S.l.: s.n.], 2024. p. 120–131. ISSN 2835-3560.

ZHU, H. et al. Data and information quality research: Its evolution and future. In: *Computing Handbook, 3rd ed.* [s.n.], 2014. Disponível em: ⟨https://api.semanticscholar.org/CorpusID:7041308⟩.

APÊNDICE A

Artigo	Pontuação
(KARRA; LASFAR, 2024)	7.0
(QIAN; REIF; KAHNG, 2024)	7.0
(KUMAR, 2024)	7.0
(CôTé et al., 2024)	7.0
(BECK, 2023)	6.5
(YU et al., 2024)	7.0
(GARRIDO-MUñOZ; MARTíNEZ-SANTIAGO; MONTEJO-RáEZ, 2023)	7.0
(MYERS et al., 2023)	7.0
(SCHWABE et al., 2024)	7.0
(LIN et al., 2024)	7.0
(SHANKAR et al., 2024)	6.5
(DU et al., 2024)	7.0
(KANG et al., 2024)	7.0
(SHEN et al., 2021)	7.0
(SUN et al., 2024)	7.0
(OU et al., 2024)	7.0
(LI et al., 2024a)	7.0
(VADAPALLI et al., 2024)	6.3
(LI et al., 2024b)	7.0
(ZHOU et al., 2024b)	7.0
(TRAN et al., 2022)	7.0
(ZHANG et al., 2024)	7.0
(LIU et al., 2024)	7.0
(JAIN et al., 2024)	7.0

Artigo	Pontuação
(ZHOU et al., 2024a)	6.5
(CHEN et al., 2024)	7.0
(JERNITE et al., 2022)	7.0
(NI et al., 2024)	7.0
(KARRA; LASFAR, 2023)	7.0
(BOJIC et al., 2023)	7.0
(WANG et al., 2024)	7.0
(ALBALAK et al., 2024)	7.0
(GUNASEKAR et al., 2023)	7.0
(LONGPRE et al., 2023)	7.0
(PENEDO et al., 2023)	7.0
(BARIKERI et al., 2021)	7.0
(FILIPPOVA, 2020)	7.0
(ROZIèRE et al., 2024)	7.0
(TOUVRON et al., 2023)	7.0