DESAFIOS DA LINGUÍSTICA DE CORPUS IMPOSTOS PELA INTELIGÊNCIA ARTIFICIAL: REDISCUTINDO ALGUNS CONCEITOS

Challenges of Corpus Linguistics Posed by Artificial Intelligence: Rethinking Some Concepts

DOI: 10.14393/LL63-v41S-2026-05

Jackson Wilke da Cruz Souza*

RESUMO: Neste artigo discuto os impactos da Inteligência Artificial (IA), especialmente dos *Large Language Models* (LLMs), sobre conceitos fundamentais da Linguística de *Corpus* (LC), a saber, autenticidade, naturalidade e representatividade. A partir de uma revisão de pesquisas recentes, evidencio a crescente integração entre IA e LC e os desafios metodológicos decorrentes, como a escolha adequada de textos e a distinção entre produções autênticas e artificiais. Ainda, discuto sobre a inclusão indiscriminada de textos gerados por IA nos *corpora* e como isso pode comprometer a validade das análises linguísticas. Isso culminará em questões técnicas e implicações simbólicas relacionadas ao conceito de autoria, identidade e confiança nos modos de circulação e produção científica. Por fim, proponho alguns encaminhamentos vislumbrando oportunidades dessa aproximação entre LC e IA.

PALAVRAS-CHAVE: Linguística de *Corpus*. Inteligência Artificial. Autenticidade. Representatividade. *Large Language Models.*

ABSTRACT: In this article, I discuss the impacts of Artificial Intelligence (AI), particularly Large Language Models (LLMs), on fundamental concepts in Corpus Linguistics (CL), namely authenticity, naturalness, and representativeness. Based on a review of recent research, I highlight the growing integration between AI and CL, as well as the resulting methodological challenges, such as the appropriate selection of texts and the distinction between authentic and artificial productions. Furthermore, I address the indiscriminate inclusion of AI-generated texts in *corpora* and how this practice may compromise the validity of linguistic analyses. These issues lead to both technical concerns and symbolic implications related to authorship, identity, and trust in scientific production and dissemination. Finally, I propose possible directions that envision opportunities in the convergence between CL and AI.

KEYWORDS: *Corpus* Linguistics. Artificial Intelligence. Authenticity. Representativeness. Large Language Models.

*Doutor em Linguística pelo Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos. Professor no Instituto de Ciência, Tecnologia e Inovação e do Programa de Pós-Graduação em Língua e Cultura da Universidade Federal da Bahia. ORCID: 0000-0003-1881-6780. E-mail: jackcruzsouza(AT)gmail.com.

1 Introdução

É notório que as redes sociais têm sido fonte dos processos de produção, circulação e recepção de conteúdos que interessam a diferentes segmentos sociais. Tais processos foram potencializados com as últimas ondas da Web (Passarelli; Gomes, 2020), fazendo com que o usuário não apenas consumisse, mas também gerasse conteúdo na internet, trazendo à tona o conceito de *User Generated Content* (UGC).

Santos (2022), ao estudar a proeminência de estudos que observam UGC, aponta que há quatro grandes temas de interesse recentes: (1) conteúdo criado pelo usuário, (2) práticas comunicativas, como o jornalismo, (3) estudos sobre usos linguísticos que ocorrem na interlocução entre usuário e audiência e (4) plataforma Web 2.0. O autor ainda destaca que UGC tem sido foco nas disciplinas de Comunicação social, mídia e jornalismo, Consumo e negócios, Ciência da informação, Ciências humanas e Ciência política.

Nas áreas de Processamento de Linguagem Natural (PLN) e Linguística de *corpus* (LC), UGC tem despertado interesse especialmente em pesquisas que visam o processamento de diferentes níveis da língua (como o morfológico, sintático e semântico), além de compreender perfis ideológicos e comportamentais de usuários a partir de suas construções textuais. É importante destacar que, nesse contexto, existem diferentes vertentes metodológicas sobre as pesquisas em LC. Há trabalhos que desenvolvem pesquisas com corpus assíncrono, em que os textos são compilados e processados fora de um ambiente on-line; já outros, com corpus síncrono, em que a compilação e o processamento se dão de maneira on-line e contínua; por fim, outros que utilizam a web como corpus, em que o material linguístico está sendo produzido e pensado de maneira síncrona ou assíncrona.

O que destaco é que em todas essas abordagens há conceitos propostos pela LC que nos serviram de base até aqui. Porém, por mudanças socio-históricas, precisamos repensar esses conceitos, especialmente porque as fronteiras entre o digital e o não digital estão cada vez mais borradas (Burnham, 2014). Ainda que discussões sobre IA em PLN não sejam novidade, elas se tornaram populares devido ao destaque midiático que se teve por conta dos *Large Language Models* (LLM). Esses modelos podem ser de língua geral, como o GPT (do inglês, *Generative Pre-trained Transformer*), ou de domínio, como o BloombergGPT (Wu *et al.*, 2023).

Em ambos os tipos, os modelos são treinados a partir de grande quantidade de dados para que processem aspectos das línguas naturais.

Assim, meu objetivo neste trabalho é discutir como conceitos caros à LC, como autenticidade, naturalidade e extensão, requerem novas discussões frente ao cenário de *Big data* e IA.

2 Trabalhos relacionados

Podemos dizer que as áreas de PLN e LC são próximas graças ao fator computacional estruturante comum a elas. Desde sua origem, a LC já incorporava recursos computacionais para viabilizar o processamento de grandes volumes de dados textuais a partir de técnicas computacionais usadas para a extração, tratamento e análise desses dados. Viana e Tagnin (2015) destacam que esse avanço se deu, inclusive, por meio de contribuições de áreas correlatas à LC, como Lexicografia e Terminologia, com abordagens direcionadas por *corpus*.

Nos últimos anos, é possível notar um crescimento significativo de estudos que integram ferramentas de IA à análise de dados linguísticos, sobretudo na abordagem *corpus-driven*. Ainda que a natureza desses estudos seja empírica, eles contribuíram para que fossem feitas provocações teóricas as quais quero retomar neste trabalho.

O trabalho de Pace-Sigge (2018) aborda a *Lexical Priming Theory* como um modelo que relaciona a análise de *corpus* à possibilidade de compreender a linguagem por meio de dados reais. O modelo teórico em questão aborda como é possível observar a combinação de palavras em determinados contextos a partir da recorrência dela no conjunto de dados. Esse trabalho foi um dos primeiros a evidenciar a interface entre IA e LC, retomando propostas algorítmicas que subsidiaram a criação de assistentes de voz, por exemplo, e a combinação do processamento simbólico amparado por estratégias estatísticas. Por outro lado, o autor já apresentava a limitação de compreensão semântica em abordagens puramente automáticas de exploração de dados, já que esse tipo de metodologia compreende o significado por semelhança contextual de co-ocorrência de palavras. Nesse cenário, os *corpora* são importantes para validar e interpretar os resultados de análises puramente automáticas em relação ao que os humanos produzem.

Curry, Baker e Brookes (2024) investigaram o uso do ChatGPT em tarefas desenvolvidas na LC, como categorização semântica de palavras-chave, análise de concordanciadores e forma-função. No trabalho, os autores destacam que, muito embora o LLM utilizado apresente resultados satisfatórios, há tendência à inferência de maneira especulativa e à alucinação do que foi gerado, comprometendo a escalabilidade da tarefa e reprodutibilidade dos dados. Nesse caso, um dos efeitos causados ao conjunto de dados seria a manipulação e/ou distorção das informações sob uma aparente performance de naturalidade discursiva, comprometendo eticamente os produtos e conclusões que seriam derivados das análises.

Consonante a essas considerações, Sardinha (2024), a partir de sua proposta multidimensional sobre *corpus*, compara produções acadêmicas feitas por humanos e por LLMs. Para tanto, o autor controla algumas variáveis linguísticas, como complexidade sintática, frequência lexical e registro pragmático. Os resultados apontam que os textos gerados por LLMs apresentam vocabulário exagerado e sofisticado do ponto de vista linguístico, além de serem redundantes e com estruturas sintáticas empobrecidas. A lição tirada desse estudo é a questão se textos gerados por IA devem ou não ser acrescidos ao *corpus*, aspecto que será retomado mais adiante.

Outra perspectiva que relaciona IA e LC é a de Cheung e Crosthwaite (2025), que propuseram um sistema para aprimorar a escrita acadêmica de estudantes. No trabalho os autores apresentam o *CorpusChat*, que é uma ferramenta que combina interfaces baseadas em corpora com modelos generativos, proporcionando ao aluno ter acesso a um banco de dados para que, tendo acesso a como redigir certas seções do texto acadêmico, possa ser letrado com base em dados artificiais e sugestões da IA. Ainda que seja bastante promissor, os autores reconhecem a necessidade de garantir que as sugestões fornecidas partam de evidências linguísticas autênticas, evitando interferências normativas ou estilísticas que podem comprometer o aprendizado do aluno.

Por fim cabe destacar a pesquisa de bastante fôlego de Shormani (2024), que realiza uma análise quantitativa de 51 anos (entre 1974 e 2024) de publicações que relacionam direta ou indiretamente Linguística e IA. Para tanto, o autor faz uma revisão da literatura utilizando as ferramentas CiteSpace (Chen, 2014) e VOSviewer (Van-Eck e Waltman, 2013) para mapear tendências temáticas e redes de (co)autoria. Dentre os resultados, o trabalho destaca que

somente em 2023 foram publicados 1478 artigos que relacionam as duas áreas de foco do estudo, evidenciando que há um emergente interesse entre LC e IA.

Apesar de os trabalhos aqui abordados majoritariamente não apresentarem reflexões teóricas sobre conceitos caros à LC, é interessante notar como eles subsidiam discussões com as quais lido na próxima seção. Além disso, essas pesquisas ajudam a corroborar como os desafios atuais e futuros que enfrentamos utilizam dados em aplicações ou abordagens de IA.

3 Revisitando conceitos

No início da LC, os *corpora* mais extensos tinham em torno de 1 milhão de palavras, como o projeto *Brown corpus* que era composto de 500 textos de 2000 palavras em 15 gêneros. Mais tarde, foram publicados outros *corpora* que tinham uma quantidade de palavras muito mais elevada, como o *corpus* multilíngue *News on the Web*, que soma mais de 5 bilhões de palavras (Tagnin, 2018).

Aqui cabe, então, discutir o primeiro conceito que quero abordar: a *representatividade*. Sinclair (1991) defende que um *corpus* representativo deva ser o maior possível, já que a linguagem é um sistema probabilístico (Halliday, 1991) e esse recurso se caracteriza por ser "uma amostra de uma população cuja dimensão não se conhece (a linguagem como um todo)" (Sardinha, 2000, p. 342). Tendo esse pressuposto, no âmbito da língua geral, o *corpus* deve ser extenso, para que, a princípio, a amostra possa ser representativa a fim de evidenciar determinados fenômenos linguísticos e/ou hipóteses de pesquisa.

Na tentativa de garantir esse critério nos *corpora*, deparamo-nos com ao menos dois desafios. O primeiro é de ordem metodológica. Sabemos que as áreas de LC e PLN dispõem de métodos bastante eficazes e robustos para a coleta e processamento de textos, como *Web Scraping* (Zhao, 2022). Nesse método, em especial, os dados são extraídos da Web e salvos em sistemas de arquivos ou banco de dados para que possam ser analisados ou recuperados posteriormente. Tal procedimento pode ser feito manualmente ou a partir de técnicas computacionais. Nesse último caso, a partir da disponibilização de um endereço eletrônico sem que haja as especificações paramétricas corretas, o sistema poderá trazer indiscriminadamente todo o conteúdo do site. Não é trivial, portanto, pontuar que os textos que são selecionados

para os *corpora* devam ser submetidos a uma curadoria em função do objetivo da pesquisa e sobre o que ele deve representar.

O segundo desafio que vislumbro diz respeito ao *processamento*. A partir da coleta dos dados, será necessário processar as informações linguísticas. Caso a coleta dos dados linguísticos seja feita por métodos puramente automáticos e não for feito nenhum tratamento para classificação de textos que foram produzidos por humanos ou por IA, as informações que serão extraídas dos dados poderão ser artificiais. Atualmente, há esforços para identificar textos gerados por IA e por humanos, como o trabalho de Ayapova e Skripnikova (2022), que classifica textos jornalísticos. Entretanto, esse tipo de processamento dos dados ainda parece distante dos estudos em LC, pois deveria estar aliado aos métodos utilizados nas pesquisas, como apontado anteriormente.

Recaímos, portanto, sobre outro conceito da LC que destaco aqui: a *autenticidade* dos conjuntos de textos. A literatura (Sinclair, 1991; Sardinha, 2000; Freitas, 2024) aponta para a necessidade de os textos que compõem o *corpus* serem autênticos. Isso significa que o material compilado deva ser produzido por humanos. Reunir textos autênticos é garantir que os resultados descritos e analisados refletem, de fato, a língua e como seus falantes a utilizam. Ao extrair informações linguísticas de seus contextos naturais de uso, a depender do objetivo, deve-se prever a compilação de textos produzidos por usuários da Web. Isso se justifica pelo fato de estarmos lidando com "produtores de conteúdo" e não apenas com "usuários de redes sociais".

Porém, diante da possibilidade de existirem produtos disponíveis que podem ser compilados sem terem sido produzidos intelectualmente por humanos, a autenticidade tornou a ter relevância. Atualmente, sabe-se sobre a existência de obras literárias completas que foram geradas a partir de IA, como discutido por Dalte (2020). Há ainda algumas discussões que giram em torno de questões éticas, como os direitos autorais da obra (Garcia, 2020).

O que levanto aqui não é o fato de que os textos artificiais não devam figurar nos *corpora*, mas sim a maneira como eles estão agregados à coletânea de textos. Cabe, então, ao pesquisador tomar dois cuidados, ao menos: (i) caso opte por métodos automáticos, deve-se aplicar métodos suplementares de identificação e classificação de textos artificiais e naturais que farão parte do conjunto de dados; (ii) caso opte por incluir textos artificiais em seu conjunto

de dados, deverá identificá-los e alertar o consulente sobre essa característica. Textos artificiais podem fazer parte do conjunto de dados linguísticos, desde que o propósito sobre esse tipo de inclusão não prejudique descrições linguísticas ou ainda tome algo como verdadeiro.

O ponto central não é necessariamente excluir os textos artificiais dos *corpora* linguísticos, mas sim questionar a forma como esses textos são integrados às bases de dados. Como dito, a inclusão indiscriminada de textos gerados por IA pode comprometer princípios fundamentais da LC. Nesse caso, portanto, é importante que o pesquisador assuma a responsabilidade metodológica de aplicar ferramentas suplementares de identificação e classificação de textos, de modo a distinguir com objetividade o que foi produzido por humanos e o que é resultado de um sistema meramente artificial.

Cabe ainda reforçar que textos artificiais podem eventualmente figurar nas pesquisas e nos *corpora* linguísticos, desde que estejam alinhados ao objetivo da investigação, sendo que os textos não serão tomados como equivalentes aos enunciados humanos. Por exemplo, investigações sobre regularidade sintática, previsibilidade lexical ou padrões de coocorrência podem se beneficiar da análise comparativa entre produções humanas e artificiais, como visto na seção anterior. A problemática em questão é quando tais textos são agregados aos *corpora* sem controle ou curadoria, gerando conclusões enviesadas, que não representam o comportamento linguístico autêntico, humano, mas apenas simulações estatísticas da linguagem.

A linguagem sempre foi compreendida não apenas como um sistema de signos, mas como expressão de subjetividade, de memória cultural e de presença humana no mundo. Em um momento em que a IA começa a ocupar esse espaço (no sentido de escrever textos acadêmicos, (re)produzir narrativas ou ainda participar de debates) estamos sendo confrontados com um novo paradigma comunicativo: "*quem fala?*" passou a ser uma pergunta ambígua. A meu ver, a autenticidade deixou de ser apenas uma questão técnica e passou a ser um problema simbólico de identidade. Afinal, quando um texto artificial reproduz com perfeição os modos de dizer de uma comunidade linguística, bem como seus usos discursivos e enunciativos, o que está em jogo é a própria noção de autoria, de pertencimento e de agenciamento do que é enunciado.

Além disso, esse processo põe em evidência o próprio rito simbólico de produção de conhecimento. A escrita científica, por exemplo, é caracterizada por uma cadeia de responsabilidades: do autor, do orientador, do revisor, do leitor crítico. Caso um LLM assuma parte ou todas as tarefas desse processo, estamos lidando com uma reconfiguração das relações de confiança não apenas de autoria, mas, sobretudo, na ciência. O texto, nesse sentido, passou a ser um espaço de incertezas quanto à origem de seu conteúdo. Isso não significa necessariamente a negação da IA, mas sim a urgência de repensar os valores simbólicos que orientam nossa relação com a linguagem.

Penso, então, que, diante desses desafios, uma possível solução esteja na criação de protocolos metodológicos híbridos, que combinem o uso de tecnologias automatizadas com a curadoria crítica humana. Isso nos obrigará a desenvolver critérios explícitos para a inclusão de textos artificiais nos *corpora*, assim como etiquetagens semânticas ou pragmáticas que sinalizem sua origem, seu propósito e seu grau de intervenção computacional. Além disso, é possível propor uma nova tipologia de *corpus*, que inclua categorias como "*corpus* misto", "*corpus* artificialmente expandido" ou "*corpus* comparativo IA-Humano". Isso poderá permitir um uso ético e objetivo dos textos gerados por algum LLM.

Por fim, talvez estejamos no início de uma nova fase teórica da LC: aquela que exige revisões conceituais sobre o que entendemos por uso real da linguagem. Termos como autenticidade, naturalidade e intenção comunicativa precisarão ser reavaliados à luz dos contextos de produção artificial. Isso não invalida o trabalho com IA, mas nos convida a (re)elaborar uma ética discursiva e epistemológica sobre o próprio dado linguístico, em que a procedência dos textos, seus modos de (re)produção e seus efeitos comunicativos sejam sempre levados em conta.

5 Considerações finais

Neste artigo meu propósito foi discutir alguns conceitos da LC sob a luz e os desafios impostos pela Inteligência Artificial aos trabalhos atuais com corpora. É quase impossível voltarmos atrás: estamos presenciando um momento em que passamos a produzir conteúdo, sobretudo na Web, e esse conteúdo autêntico disputa espaço, muitas vezes, com um conteúdo artificial. O que propus aqui, então, foi uma breve reflexão sobre como esses conteúdos podem

e devem figurar em nossos estudos de descrição e de aplicação da linguagem frente aos métodos e ferramentas computacionais.

Há muitos outros desafios que devem ser debatidos nesse campo e muitos outros conceitos da própria LC que merecem ser discutidos, que culminarão em uma nova tipologia e organização e compreensão do objeto *corpus*. Alguns desses desafios já foram enfrentados em outras disciplinas próximas, como a Linguística Aplicada. Porém, há outros que ainda estão surgindo por conta do avanço da IA; para esses, as soluções ainda estão sendo pensadas conforme os desafios surgem.

Por fim, pondero que as pesquisas em LC não se findaram por estarmos desafiados. Muito pelo contrário: temos novos caminhos a trilhar, o que traz novo fôlego à área e outras perspectivas de pesquisa. Minha contribuição aqui é chamar a atenção para nossas metodologias de pesquisa e as concepções que estamos utilizando sobre determinados conceitos cunhados há algum tempo, antes da evolução da IA. Importa dizer, então, que o mundo que conhecemos pela literatura clássica em LC mudou por conta do computador; nossos conceitos sobre ela, então, também devem ser repensados. Ao invés de substituir os fundamentos da LC, é possível que a IA possa servir de propulsão para novas reflexões sobre os limites e as possibilidades da linguagem nessa nossa era digital.

Agradecimentos

Este trabalho foi apoiado por recursos financeiros da Chamada PRPPG-UFBA 010/2024 – Programa de Apoio a Docentes/Pesquisadores em Início de Carreira 2024 – e da Chamada FAPESB/CNPq 004/2023 – Programa Primeiros Projetos.

Referências

AYAPOVA, S. M.; SKRIPNIKOVA, A. I. Ai and human created media texts: experiment results. **Herald of journalism**, v. 64, n. 2, 2022. DOI: https://doi.org/10.26577/hj.2022.v64.i2.08

BURNHAM, T.F. Reconstrução-síntese de contribuições ao XI CINFORM, à guisa de apresentação. In BORGES, J; BARREIRA, M.I.J.S.; CUNHA, F.J.A.P. (Org.). **Mundo digital**: uma sociedade sem fronteiras? 1ed.João Pessoa: Ideia, 2014, v. 1, p. 7-20.

CURRY, N.; BAKER, P.; BROOKES, G. Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. **Applied Corpus Linguistics**, v. 4, n. 1, p. 100082, 2024. DOI: https://doi.org/10.1016/j.acorp.2023.100082

CHEN, C. The citespace manual. **College of Computing and Informatics**, v. 1, n. 1, p. 1-84, 2014.

CHEUNG, L.; CROSTHWAITE, P. CorpusChat: integrating corpus linguistics and generative AI for academic writing development. **Computer Assisted Language Learning**, p. 1–27, 2025. DOI: https://doi.org/10.1080/09588221.2025.2506480.

DALTE, P. Inteligência artificial e poesia. **Revista 2i: Estudos de Identidade e Intermedialidade**, v. 2, n. 2, p. 165–177, 2020. DOI: https://doi.org/10.21814/2i.2505

FREITAS,. Dataset e corpus. *In*: CASELI, Helena de Medeiros; NUNES, Maria das Graças Volpe (Orgs.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. [s.l.]: BPLN - Brasileiras em PLN, 2023, p. 1–37. Disponível em: <https://brasileiraspln.com/livro-pln/2a-edicao/parte-dados-avaliacao/cap-dataset-corpus/cap-dataset-corpus.pdf>.

GARCIA, A.C. Ética e inteligência artificial. **Computação Brasil**, n. 43, p. 14-22, 2020. https://doi.org/10.5753/compbr.2020.43.1791.

HALLIDAY, M.A.K. Corpus studies and probabilistic grammar. In AIJMER, K.; ALTENBERG, B. (Orgs). **English corpus linguistics**: Studies in honour of Jan Svartvik, London: Longman, 2014. p. 42-55.

PACE-SIGGE, M. Where Corpus Linguistics and Artificial Intelligence (AI) Meet. In: **Spreading Activation, Lexical Priming and the Semantic Web**: Early Psycholinguistic Theories, Corpus Linguistics and AI Applications. Cham: Palgrave Pivot, 2018. p. 29–82. DOI: 10.1007/978-3-319-90719-2_3.

PASSARELLI, B.; GOMES, A.C.F. Transliteracias: A Terceira Onda Informacional nas Humanidades Digitais. **Revista Ibero-Americana de Ciência da Informação**, v. 13, n. 1, p. 253–275, 2020. DOI: https://doi.org/10.26512/rici.v13.n1.2020.29527

SANTOS, M.L.B. The “so-called” UGC: an updated definition of user-generated content in the age of social media. **Online Information Review**, v. 46, n. 1, p. 95–113, 2021. DOI: https://doi.org/10.1108/oir-06-2020-0258

SARDINHA, T.B. AI-generated vs human-authored texts: A multidimensional comparison. **Applied Corpus Linguistics**, v. 4, n. 1, p. 100083, 2024. DOI: https://doi.org/10.1016/j.acorp.2023.100083

SARDINHA, T.B. Linguística de Corpus: histórico e problemática. **DELTA: Documentação de Estudos em Linguística Teórica e Aplicada**, v. 16, n. 2, p. 323–367, 2000. DOI: https://doi.org/10.1590/s0102-44502000000200005

SINCLAIR, J. **Corpus, Concordance, Collocation**. [s.l.]: Oxford University Press, USA, 1991.

SHORMANI, M.Q. What fifty-one years of Linguistics and Artificial Intelligence research tell us about their correlation: A scientometric review. **arXiv:2411.19858**, 2024. DOI: https://doi.org/10.48550/arXiv.2411.19858

TAGNIN, S.E.O. A Linguística de Corpus vai desbravando novos horizontes. In FINATTO, MJB; REBECHI, T.; SARMENTO, S; BOCORNY, A. EP (Org). Linguística de corpus: perspectivas. Porto Alegre: Instituto de Letras da UFGRS, p. 11-15, 2018.

VAN ECK, N.Jan; WALTMAN, L. VOSviewer manual. **Leiden: Univeristeit Leiden**, v. 1, n. 1, p. 1-53, 2013.

VIANA, V.; TAGNIN, S.E.O. Corpora na Tradução. São Paulo: HUB Editorial, 2015.

WU, S.; IRSOY, O.; LU, S.; *et al*. **BloombergGPT: A Large Language Model for Finance**. arXiv.org. Disponível em: <https://arxiv.org/abs/2303.17564>.

ZHAO, B. Web Scraping. *In*: **Encyclopedia of Big Data**. Cham: Springer International Publishing, 2022, p. 951–953. DOI: http://dx.doi.org/10.1007/978-3-319-32010-6_483.

Recebido em: 14.08.2026 Aprovado em: 31.03.2026