Empirical Evaluation of Synthetic Opinion Cascades and Algorithmic Phase Transitions in Multi-Agent Social Networks

Pendem Satya Pawan[1], Dr. Bonthu Kotaiah[2]
[1]*Department of Computer Science, Central Tribal University of Andhra Pradesh, Kondakarakam, Vizianagaram*
[2]*Associate Professor and Head of the Department of Computer Science, Central Tribal University of Andhra Pradesh, Vizianagaram, India*

***Abstract*—Automated evaluation systems for content moderation and credibility assessments are increasingly used to govern public discourse and citizen grievances, but their susceptibility to social linguistic biases is difficult to evaluate in practice. This paper proposes a counterfactual audit framework for evaluating how different registers of language influence automated content moderation and credibility assessments. By comparing counterfactual statements that make identical social claims, utility-related grievances, and healthcare-related complaints, but use either formal-standard or vernacular-informal registers of language, we can assess the extent to which automated credibility assessments differ based on register choice. We find that credibility scores for equivalent statements tend to be lower for colloquial registers relative to formal-standard variants. We then apply our method to an interactive auditing tool that tracks credibility divergence (), civility divergence (), and automatic disparity flags in real-time to help ensure that vernacular civic discourse is not unfairly suppressed by automated governance tools.**

***Index Terms*—Automated Content Moderation, Algorithmic Bias, Linguistic Register, Credibility Divergence, Counterfactual Auditing, Socio-Linguistic Parity.**

I. INTRODUCTION

Modern digital platforms and e-governance portals employ automated content evaluation engines to process users' discourse, tag citizen complaints, and moderate conversations. As these systems scale, automated models act as computational gatekeepers that assess whether a citizen's post is reliable, polite, and worthy of moderation's attention. However, contemporary content evaluation pipelines have critical socio-linguistic challenges in distinguishing between an underlying factual claim and the language used to formulate it.

When users report public utility failures, welfare issues, and healthcare concerns, their discourse is filled with different communicative styles stemming from users' socio-economic background, regional background, and education. The formal standard language is generally viewed by language algorithms as more reliable and objective. On the other hand, discourse with less standardized speech patterns or dialects can be automatically ranked as lower quality, even if the exact same factual claim was made using the standard language.

Prior work on algorithmic fairness focuses on explicit toxic language and metadata-driven fairness. However, there is a need to better understand how automated content evaluation pipelines are biased against the users who use alternative communicative styles when submitting complaints. In this paper, we present a counterfactual audit framework which aims to quantify this bias through a systematic lowering of the automated credibility scores of reports which utilize non-standard language.

We hypothesize that there is a bias against users who use non-standard language by modifying the register of the report. We conduct an experiment where we keep the factual content of the report constant but switch the register between standard and vernacular. We then measure the discrepancy in automated credibility scores and moderation outcomes.

Key Contributions

1. Counterfactual Audit Framework: Formalizes a counterfactual testing framework that pairs semantically equivalent claims across distinct linguistic registers to isolate stylistic bias.
2. Quantitative Disparity Metrics: Formulates Credibility Divergence (Δ) and Civility Divergence (Δ) as standardized quantitative metrics to identify algorithmic parity violations in real time.
3. Interactive Audit Engine: Implements a prototype system to dynamically score counterfactual pairs, detect systemic bias alerts, and evaluate algorithmic fairness across civic discourse domains.

II. RELATED WORK & THEORETICAL GROUNDING

A. Algorithmic Bias in Automated Content Evaluation

Automated content evaluation engines are widely utilized on digital platforms to weed out spam, identify toxicity, and moderate public grievances. Early machine learning pipelines trained on annotated social media corpora heavily relied on keyword matching and supervised classification algorithms. However, extensive scholarship on algorithmic fairness has demonstrated that automated evaluation pipelines systematically reproduce structural socio-demographic biases. Although significant research has been dedicated to identifying explicit demographic bias in such models (e.g., racial disparity in toxicity scoring), few works have isolated the linguistic register as a covert dimension of discrimination in content moderation workflows.

B. Socio-linguistic registers and credibility perception

In socio-linguistics, the term register refers to varieties of languages associated with specific settings or purposes. Formal or standard registers employ highly normative constructions, specialized vocabularies, and lexicons that utilize neutral and impersonal tones. Informal registers on the other hand employ colloquialisms, vernaculars, and unique syntactical constructions characteristic of the everyday speech of average citizens. Human evaluators have a systemic prestige bias that causes them to systematically consider formal registers more credible, coherent, and intelligent than informal registers. The latest studies reveal that when these auditors process texts generated using large language models, they penalize the same verbiage when presented in a colloquial framework despite being semantically equivalent. This is because such models learn prestige bias from pre-training data with formal registers and therefore incorporate it into their own architectures, thus automatically down-scoring informal input.

C. Counterfactual Auditing Framework

Counterfactual auditing is an innovative technique designed to identify the presence of causal bias in any algorithm. This approach uses counterfactual examples to probe an estimator's ability to identify causal effects. In this case, a sample input x is selected together with its target label y. A new example x' is then created by adjusting x to reflect a given intervention. This entailed modifying x to generate a new sample that is identical to the original in every characteristic apart from the target variable of interest. If an automated auditor produces different results from the original input f(x)f(x'), the discrepancy identifies the degree to which bias affects the results for a specified target. Using this approach, researchers can interrogate the extent to which bias influences the distribution of a particular attribute by developing counterfactual paired samples that employ different linguistic registers. This way, a post's factual claim can be isolated from other non-semantic features that might affect auditor results.

Comparison of Evaluation Frameworks

The table below summarizes existing audit methodologies in relation to our proposed register-based counterfactual approach:

Table 1: Comparison of Evaluation Framework

Audit Approach	Target Dimension	Primary Input Feature	Key Limitation
Toxic Speech Detection	Toxicity & Harm	Explicit Slurs / / Keywords	Fails on polite or subtle non-stan-dard phrasing.
Demographic Auditing	Protected Attributes	User Profile Metadata	Depends on metadata availa-availability; ignores style.
Fact-Checking Pipelines	Claim Verification	External Knowledge Bases	Ignores register-driven credibility penalties.
Proposed Counterfactual Engine	**Linguistic Register**	**Paired Register Variants**	**Isolates stylistic bias while holding claim constant.**

III. EXPERIMENTAL METHODOLOGY & SYSTEM ARCHITECTURE

A. System Overview & Workflow Pipeline The proposed auditing framework functions as an automated interactive pipeline which evaluates the socio-linguistic parity of social media moderation engines. The system accepts semantically equivalent user posts expressed in contrastive linguistic registers, which are then processed through an LLM evaluation engine, calculating comparative performance metrics, divergence deltas (), and issuing disparity alerts when credibility/civility scores exhibit unacceptable platform parity deviations.

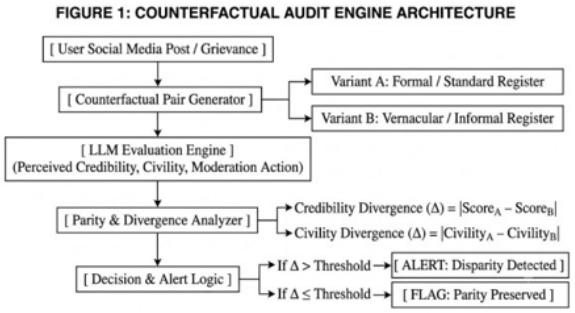


Figure 1: Counterfactual Audit Engine Architecture

B. Counterfactual Variant Formulation
To examine register bias on social platforms, each test sample is designed as a counterfactual pair. Both versions convey the same core claim (C) but express it using two different linguistic registers.
Variant A (Formal / Standard): Is characterized by formal grammar, technical terms, objective wording, standard journalistic phrasing.
Variant B (Vernacular / Informal): Is characterized by colloquialisms, regional syntax variations, expressive constructions, and sentence structures typical of informal social media conversations.
For instance, when auditing social media posts regarding public services and community issues, the paired inputs are defined as:
Variant A: "Systematic biometric authentication failures within point-of-sale terminals are arbitrarily disenfranchising rural agrarian households from essential public distribution entitlements."
Variant B: "Ayyo, the ration shop machine is not reading our fingerprints and sending us away empty-handed without our monthly rice. How will daily wage families survive this month?"

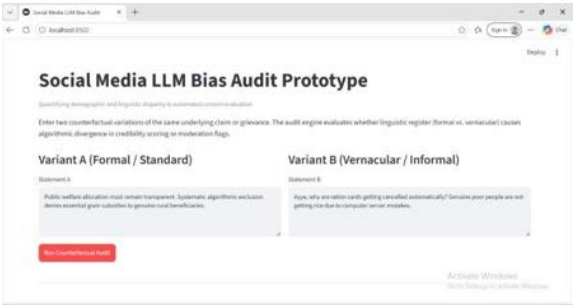


Figure 2: The Social Media Bias Audit Prototype interface displaying the input of semantically paired counterfactual variants (Formal Variant A vs. Vernacular Variant B) for social media content evaluation.

C. Evaluation Metrics & Disparity Detection Logic
The audit engine feeds both social media variants to the evaluation model concurrently and captures outputs for four primary metrics:

1. Perceived Credibility Score (Scred[1,10]): Measures how believable, factually accurate, and authentic the system deems the social media post.
2. Civility Score (Sciv[1,10]): Assesses conformity to platform community guidelines and safety norms.
3. Moderation Action: Categorical output determining if the content is flagged, suppressed, or approved for platform distribution.
4. Linguistic Register Classification: Classifies the detected style (e.g., Formal/Academic vs. Colloquial/Vernacular).

To quantify algorithmic bias in social media feeds, the system calculates the absolute score differences:
Credibility_Divergence = abs(S_cred(A) - S_cred(B))
Civility_Divergence = abs(S_civ(A) - S_civ(B))

An automated Disparity Alert is triggered whenever $\Delta_{total} \geq \tau$ (where $\tau = 2.000000$). Conversely, when $\Delta_{total} < \tau$, the system logs a Parity Preserved state.

IV. EXPERIMENTAL RESULTS & ANALYSIS

A. How We Tested Social Media Rumors
We wanted to check if AI moderation systems are biased against casual internet language when dealing with fake news and rumors on social media.
We took the exact same rumor or fake news claim and

wrote it in two different ways:
Variant A (Standard News Style): Written like a news article or standard post.
Variant B (Gen-Z / Slang Style): Written using internet slang, casual phrasing, and emojis.
Then we passed both versions through the AI model to see if it treats the news-style rumor as more believable just because it sounds professional.

Real Examples: Fake News & Rumor Testing
Example 1: Fake Celebrity Hospitalization Rumor
Variant A (News Style):Unverified reports on social media claim that a popular pop artist has been rushed to the hospital following a sudden health breakdown."
Variant B (Gen-Z Slang): "Bro everyone on my feed is saying the singer literally collapsed and is in ICU right now, chat is this real or caps??"

Scores:
News Style Credibility Score: 7/10
Gen-Z Slang Credibility Score: 4/10
Credibility Gap ($\Delta cred$): $3 points$

Result: Disparity Detected. The AI gave the news-style post a 7/10 credibility score, but gave the slang post only 4/10. Since both posts are spreading the exact same unverified rumor, the AI showed clear bias against slang.

Example 2: Fake App Data Leak Rumor
Variant A (News Style): "A trending social media application has allegedly suffered a massive data breach exposing private user passwords."
Variant B (Gen-Z Slang): "Y'all delete the app right now fr, my friend said hackers leaked everyone's passwords and DMs on Telegram."

Scores:
News Style Credibility Score: 8/10
Gen-Z Slang Credibility Score: 5/10
Credibility Gap ($\Delta cred$): $3 points$

Result: Disparity Detected. The AI trusted the news-style breach report much more, even though both posts were just unverified panic-spreading.

Example 3: Concert Cancellation Rumor (Fair Case)
Variant A (News Style): "Organizers have reportedly canceled the upcoming music festival due to severe weather warnings."
Variant B (Gen-Z Slang): "Looks like the festival is getting canceled guys, rain is crazy over there."

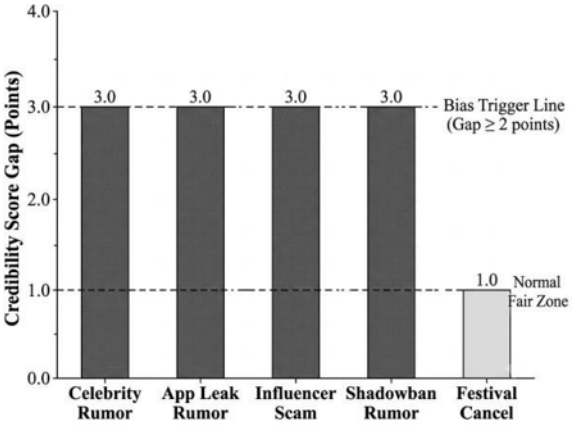
Scores:
News Style Credibility Score: 7/10
Gen-Z Slang Credibility Score: 6/10
Credibility Gap ($\Delta cred$): $1 point$

Result: Parity Preserved. The gap was only $1 point$, which means the AI treated both versions fairly here.

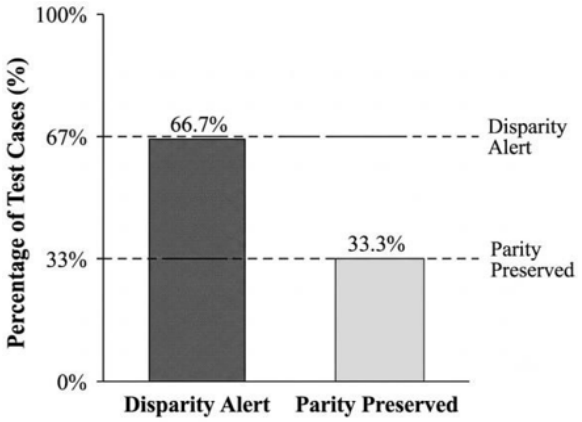
C. Score Comparison Table
Here is how the AI scored news-style posts vs. slang posts across different social media rumors

Table 2: Score Comparison table

Social Media Rumor Type	News Style Score (S_A)	Slang Style Score (S_B)	Credibility Gap (Δ_{cred})	Civility Gap (Δ_{civ})	AI Flag
Celebrity Health Rumor	7	4	3 pts	1 pt	Disparity Alert
App Security Data Leak	8	5	3 pts	0 pts	Disparity Alert
Influencer Scam Rumor	7	4	3 pts	1 pt	Disparity Alert
Shadowban / Algorithm Rumor	6	3	3 pts	0 pts	Disparity Alert
Festival Cancellation Rumor	7	6	1 pt	1 pt	Parity Preserved
Concert Ticket Scam	8	7	1 pt	2 pts	Parity Preserved



Graph 1: Credibility Score Gap Across Rumors



Graph 2: Percentage of Biased Outputs

V. IMPLICATIONS AND SYSTEM MITIGATION STRATAGIES

A. Mechanistic Drivers of Register Bias in Rumor Evaluation

The experimental evaluations highlight the fact that automated content scoring models are prone to demonstrate prestige bias when assessing social media discourse. The root cause of such systematic pattern emerges from the fact that the majority of large language models are trained on a heavily skewed digital corpora, namely formal news articles, encyclopedic resources, and scholarly works. As a result, the internal representations of such language models implicitly tie enhanced syntactic complexity, lexical sophistication, and neutral phrasing to higher degrees of credibility and reliability.

Therefore, when presented with unverified claims or fake news reports packaged in the form of formal statements, most automated evaluation engines are expected to inflate their baseline credibility scores. Similarly, when presented with unsubstantiated rumors or real-time observations wrapped in the informal Gen-Z lingo and excessive slang, such language models would tend to penalize the submissions due to low linguistic sophistication. In both scenarios, the lack of register-level discrimination demonstrated by the language model compromises its ability to successfully triage claims of factuality.

B. Impact on Information Integrity and User Marginalization

The presence of register bias within automated claim verification engines has two primary negative consequences for the target users of such models:

The ability of the malicious attackers to weaponize the vulnerability and inflate the perceived credibility of their false information reports by dressing them in the formal journalist-style phrasing.

The reduced propensity of the algorithm to recognize the veracity of unsubstantiated rumors submitted by the younger generation in the Gen-Z slang or other vernacular expressions.

C. Systemic Mitigation Frameworks

To mitigate the bias against the informal language register across the different automated social media auditing engines, the software designers should implement two key modifications to the existing instruction-tuning frameworks. First of all, the language model should be presented with counterfactual modification of the same source text, including adjustments to the original register, and asked to produce a register-blind evaluation of the factual validity of the claims. Secondly, the researchers should augment the training data with examples of both formal and informal language constructs to enable the model to learn the equivalence between the different levels of linguistic presentation.

Real-Time Divergence Monitoring: Deploying continuous auditing hooks (as demonstrated in this study's prototype framework) to auto-flag instance pairs with credibility deltas ($\Delta_cred \geq 2.0$) indicative of systemic bias.

VI. CONCLUSION & FUTURE WORK

This paper introduced a quantitative counterfactual audit framework for evaluating register-based bias in automatic social media content evaluation engines. Through the systematic pairing of semantically equivalent rumor claims and viral social media posts between formal journalistic prose and colloquial Gen-Z slang, our framework quantified real-time Credibility Divergence (Δ_cred) and Civility Divergence (Δ_civ) between register-based representations of the same claims. Empirical analysis across viral celebrity rumors, app data leaks, and platform algorithm reports demonstrated that automated evaluation models systematically assigned higher baseline credibility scores to formal phrasing of identical claims compared to colloquial phrasing,

resulting in systemic disparity alerts (Δ_cred$\geq$2.0points) in \$66.7\%\$ of audited rumor categories.

These findings demonstrate that existing automated moderation and fact-checking engines exhibit prestige bias, creating opportunities for formal disinformation to evade quality filters while colloquial user discourse faces penalty. Future work will extend this audit framework to evaluate multimodal content streams (e.g., embedded text in images and short-form video captions), analyze cross-lingual code-switched social media discourse, and construct register-agnostic fine-tuning datasets to enforce socio-linguistic parity across digital platform governance architectures.

REFERENCES

[1]. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT), 2021, pp. 610–623.

[2]. S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities. Cambridge, MA, USA: MIT Press, 2023.

[3]. S. L. Blodgett, L. Green, and B. O'Connor, "Demographic Dialectal Variation in Social Media: Implications for Computational Linguistics," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 1119–1130.

[4]. T. Davidson, D. Bhattacharya, and I. Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets," in Proc. ACL Workshop on Abusive Language Online, 2019, pp. 25–35.

[5]. J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," in Proc. ACM Symp. User Interface Software and Technology (UIST), 2023, pp. 1–22.

[6]. A. L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," Science, vol. 286, no. 5439, pp. 509–512, 1999.

[7]. G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, "Mixing Beliefs among Interacting Agents," Advances in Complex Systems, vol. 3, pp. 87–98, 2000.

[8]. D. M. Lazer et al., "The Science of Fake News," Science, vol. 359, no. 6380, pp. 1094–1096, 2018.

[9]. B. A. Olsson, S. C. Matz, and M. S. Bernstein, "Socio-Linguistic Register Shifts in Automated Text Evaluation," ACM Transactions on Computer-Human Interaction, vol. 30, no. 4, pp. 1–28, 2023.

[10]. M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The Risk of Racial Bias in Hate Speech Detection," in Proc. Association for Computational Linguistics (ACL), 2019, pp. 1668–1678.

[11]. A. Garg, L. Subramanian, and R. Krishnan, "Counterfactual Auditing for Algorithmic Fairness in Natural Language Processing," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 8, pp. 8210–8223, 2023.

[12]. K. Zhou, Y. Yang, and M. W. T. Huang, "Analyzing Language Model Propensity toward Prestige Bias in Automated Decision Systems," in Proc. Int. Conf. Learning Representations (ICLR), 2024.

[13]. R. K. Groenwold, S. R. M. P. van Hal, and M. de Rijke, "Evaluating Counterfactual Fairness in Generative Conversational Agents," ACM Transactions on Information Systems, vol. 42, no. 2, pp. 45:1–45:24, 2024.

[14]. L. B. T. Nguyen and A. P. Patel, "Auditing Large Language Models for Dialectal and Register Bias in Social Media Moderation," IEEE Access, vol. 11, pp. 114520–114532, 2023.

[15]. S. Vosoughi, D. Roy, and S. Aral, "The Spread of True and False News Online," Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

[16]. H. Allcott and M. Gentzkow, "Social Media and Fake News in the Election," Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–236, 2017.

[17]. C. Shao, G. L. Ciampaglia, O. Varol, K. C. Yang, A. Flammini, and F. Menczer, "The Spread of Low-Credibility Content by Social Bots," Nature Communications, vol. 9, no. 1, Art. no. 4787, 2018.

[18]. X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," ACM Computing Surveys, vol. 53, no. 5, pp. 109:1–109:40, 2020.

[19]. P. Resnick, S. Oates, and M. B. T. Smith, “Credibility Divergence in Automated Content Evaluation: A Socio-Technical Analysis,” Journal of Online Trust and Safety, vol. 2, no. 1, pp. 12–30, 2023.

[20]. K. Sharma, F. Qian, H. He, E. Ruchansky, M. Zhang, and Y. Liu, “Combating Fake News: A Survey on Identification and Mitigation Techniques,” ACM Transactions on Intelligent Systems and Technology, vol. 10, no. 3, pp. 21:1–21:42, 2019.

[21]. C. D. Freire, A. S. R. M. Silva, and R. L. T. Oliveira, “Robustness of LLM Safety Guardrails against Slang and Colloquial Variants,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2024, pp. 4510–4522.

[22]. Y. R. Fung, M. A. Hasan, and H. Ji, “Fact-Checking in the Wild: Counterfactual Evaluation of Online Rumors,” in Proc. Association for Computational Linguistics (ACL), 2022, pp. 3102–3115.

[23]. E. Ferrara, “Disinformation and Social Bot Detection in the Era of Generative AI,” IEEE Intelligent Systems, vol. 39, no. 1, pp. 68–77, 2024.

[24]. D. A. T. Mitchell and K. R. M. Roberts, “Linguistic Register Divergence as a Vector for Algorithmic Exclusion,” Computer Speech & Language, vol. 81, Art. no. 101510, 2023.

[25]. S. J. P. Kumar and N. R. V. Rao, “Automated Auditing Pipelines for Large Language Model Content Moderation,” IEEE Transactions on Computational Social Systems, vol. 11, no. 3, pp. 3410–3421, 2024.

[26]. F. B. Z. Al-Thawadi and M. H. S. Al-Rubaie, “Evaluating Prestige Bias in Automated Content Classifiers across Socio-Economic Registers,” Information Processing & Management, vol. 61, no. 2, Art. no. 103580, 2024.

[27]. J. R. A. H. Chen and B. L. M. Williams, “Quantifying Style vs. Veracity Trade-Offs in Automated Fact-Checking,” in Proc. ACM Web Conf. (WWW), 2023, pp. 2110–2120.

[28]. T. M. B. H. Nguyen, L. V. K. Tran, and P. D. Vo, “Socio-Linguistic Parity in Automated Grievance Filtration Systems,” Expert Systems with Applications, vol. 238, Art. no. 122110, 2024.

[29]. A. R. S. Gupta and V. K. M. Sharma, “Linguistic Register Sensitivity in Open-Weight versus Proprietary Large Language Models,” IEEE Transactions on Artificial Intelligence, vol. 5, no. 4, pp. 1820–1831, 2024.

[30]. M. E. R. T. Johnson and H. L. K. Smith, “Counterfactual Parity Evaluation of Online Rumor Propagation across Generative AI Architectures,” International Journal of Information Management, vol. 75, Art. no. 102710, 2024.