CrediBench: Building Web-Scale Network Datasets for Information Integrity

Emma Kondrup[1,2]* **Sebastian Sabry**[1,2]* **Hussein Abdallah**[1] **Zachary Yang**[1,2]
James Zhou[6] **Kellin Pelrine**[1,2] **Jean-Françis Godbout**[1,4]
Michael M. Bronstein[3,7] **Reihaneh Rabbany**[1,2,5] **Shenyang Huang**[3]
[1]Mila - Quebec AI Institute, [2]School of Computer Science, McGill University,
[3]University of Oxford, Oxford, UK, [4]Université de Montréal, [5]CIFAR AI Chair,
[6]UC Berkeley, California, USA, [7]AITHYRA,Vienna, Austria

Abstract

Online misinformation poses an escalating threat, amplified by the Internet's open nature and increasingly capable LLMs that generate persuasive yet deceptive content. Existing misinformation detection methods typically focus on either textual content or network structure in isolation, failing to leverage the rich, dynamic interplay between website content and hyperlink relationships that characterizes real-world misinformation ecosystems. We introduce CrediBench: a large-scale data processing pipeline for constructing temporal web graphs that jointly model textual content and hyperlink structure for misinformation detection. Unlike prior work, our approach captures the dynamic evolution of general misinformation domains, including changes in both content and inter-site references over time. Our processed one-month snapshot extracted from the Common Crawl archive in December 2024 contains 45 million nodes and 1 billion edges, representing the largest web graph dataset made publicly available for misinformation research to date. From our experiments on this graph snapshot, we demonstrate the strength of both structural and webpage content signals for learning credibility scores, which measure source reliability. The pipeline and experimentation code are all available here, and the dataset is made publicly available on huggingface.

1 Introduction

The digital information landscape is evolving at an unprecedented pace. While the open nature of the Internet has democratized access to information, it has also amplified the scale, speed and reach of misinformation propagation. This phenomenon not only distorts public discourse but can also undermine democratic processes, public health, and social cohesion [22, 32, 38]. In fact, experts rated misinformation and disinformation as the number one imminent risk the world faces in 2024-25 [19]—urgently calling for the development of systems to ensure information veracity. The rise of generative AI further compounds this challenge: Large Language Models (LLMs) can now produce text that is increasingly indistinguishable from human-generated content [13], making misinformation more persuasive and far harder to detect at the user level.

Fortunately, AI also stands as a key path to a solution, making large-scale, advanced automated assessments possible. However, existing approaches fail to preserve all aspects of web data that factor into misinformation detection—the textual content of websites, their diverse nature and relations, and their temporal and structural relations—especially at scale. Relatedly, proposed solutions have been recognized to lack robust generalizability (to out-of-distribution data, but also to new topics and emerging information generation techniques) [26, 47]. These methods are often catered to the

*Equal contributions

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: New Perspectives in Graph Machine Learning, NeurIPS 2025.

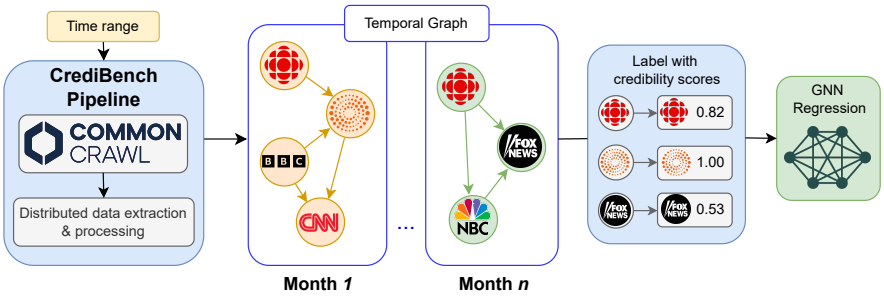


Figure 1: Overview of the CrediBench pipeline. Common Crawl data is extracted and processed to produce temporal web graphs, where nodes are web domains and edges hyperlinks. These temporal graphs, annotated with human-evaluated credibility scores, are passed to Graph Neural Networks (GNNs) for them to learn the regression task of assigning credibility scores.

small scale of existing datasets, which usually span over a limited time window or a single platform [41, 43, 34, 25, 51, 36]. These datasets fail to recognize the comprehensive, general-scoped and open nature of the Internet and how it affects information propagation. Thus, the weight of this issue on the real world calls for urgent principled systems to automate credibility assessments online at scale, which would significantly enhance the capacity to detect misinformation.

In this work, we introduce the CrediBench framework which enables an automated pipeline for assigning credibility scores to web domains, quantifying their degree of misinformation, by providing large-scale temporal text-attributed graphs that preserve both the structural and content-rich characteristics of online information sources. The temporal graph curated by CrediBench is large-scale, evolving and contains text content for its nodes, thus providing ample opportunities to apply both natural language processing techniques as well as graph ML methods for credibility score assignment to measure source reliability. Our main contributions are as follows:

- **Graph construction pipeline.** We introduce a pipeline to retrieve, decompress and construct large-scale web graphs, as illustrated in Figure 1. The pipeline builds data from Common Crawl [17], a data provider that has been crawling and making available large-scale web data on a monthly basis. Our pipeline holistically preserves the structural, temporal, and semantic facets of web data, by constructing rich temporal graphs—up to 150 million nodes and more than a billion edges per snapshot—while extracting domain-level textual and structural signals. The textual content can be embedded as node features, enabling principled graph-based learning for web domain credibility assessment.
- **Large-scale, general-scoped web graph.** We present a one-month snapshot obtained through our pipeline, specifically from December 2024. This graph has upwards of a billion edges and is annotated with human-generated credibility labels. This is one of the largest graph datasets available to date, making it an ideal foundation for future research on the credibility of web domains. Given this dataset's rich attributes and reflective real-world nature, we hope this addresses the lack of benchmark datasets in real-world applications [6].
- **Structure and text are strong signals for learning credibility.** Through experiments with both GNNs and embedded text domain content, we show that both relational and text context of the web domains provide strong signals for assigning their credibility score. This demonstrates the importance of our proposed CrediBench and the need to model both structural and textual features of web domains for credibility assignment.

Reproducibility: The code for the pipeline and experiments is made available at `https://github.com/ekmpa/CrediGraph`. The dataset is made publicly available at `https://huggingface.co/datasets/ekmpa/CrediBench`.

2 Related work

Traditional Misinformation Detection Efforts to assess the credibility of online information sources have ranged from manual, expert-driven evaluations to computational methods. Manual approaches

include user-focused frameworks such as the CRAAP test [7] or the SIFT framework [12]. Expert knowledge is also often leveraged, and most datasets for credibility on the web rely on such human annotation [37]. However, such manual methods are inherently unable to scale to modern information flows. To address this, machine and deep learning methods quickly emerged as a strong solution. These efforts have largely focused on credibility assessments at the claim level, using endogenous signals from the text, such as linguistic and stylistic features, semantic content, and multimodal cues [54, 15, 30, 46, 34]. These text features have been roughly categorized into *stylistic* features, *complexity* features and *psychological* features, all tied to disinformation detection [13]. However, these signals fail to model the complex web structure of misinformation websites and information propagation. As such, exogenous signals, such as social endorsement cues, have also been integrated in some of these works [11, 25, 43, 52, 41], but to a much lesser extent. Moreover, such works focusing on the claim-level inherently ignore the more general nature of the issue: misinformation spreads across sources and platforms, and detecting whether a certain piece of text contains misinformation largely benefits from assessing its source and other online entities' interactions with it.

LLM Based Misinformation Detection Approaches to credibility assessments online are increasingly exploring the role of LLMs for misinformation detection [14, 44, 29]. Particularly, Retrieval-Augmented Generation (RAG) has shown promise for this, as introduced in [49] which uses RAG agents for evidence-based misinformation detection. RAG indeed renders LLMs more factual and up-to-date, both crucial for detecting misinformation [49, 8, 40]. However, existing works fail to preserve all aspects of web data that factor into information veracity; namely content, structural and temporal patterns. Many are based on content, often at the claim level again, extracting features from the text. Some include social context, which can be divided into social engagement and social networks [13]; but these are weaker in the relational information they contain than entire, large-scale text-attributed web graphs.

Graph Based Methods Considering the importance of content, structural and temporal patterns, finding a principled method to incorporate all three is very compelling. Using graphs enables the modelling of complex relations between information sources and the propagation of credibility [39], as information online does not exist in isolation. For example, these approaches often leverage graph-structured data, sometimes building on link analysis algorithms [9] and using them as features to learn from [42, 42], or using graph knowledge bases [28]. Most use modern GNNs to capture propagation patterns of credibility signals through message passing [48, 55, 36]. For instance, [55] propose a dual-channel graph attention mechanism that jointly captures link structure and propagation dynamics, and [45] integrates heterogeneous graph data and adversarial active learning to better handle the diversity and noise present in online information ecosystems. While temporal graphs are particularly compelling considering they can model the dynamic nature of information propagation, their use for this task remains largely unexplored.

3 Preliminaries

We first introduce the graph notations for the temporal text-attributed graph constructed from the CrediBench pipeline. As the Common Crawl data is released monthly, we represent the graph as a sequence of graph snapshots, similar to the Discrete-time Dynamic Graph setup defined in [31]. The data formulation is as follows:

Definition 1 (Temporal text-Attributed Graphs) *A Temporal text-Attributed Graph (TAG)* $\mathbf{G}$ *is a sequence of graph snapshots sampled at regularly-spaced time intervals and nodes have evolving text content at each snapshot:*

$$\mathbf{G} = \{\mathbf{G}_0, \mathbf{G}_1, \ldots, \mathbf{G}_T\}$$

$\mathbf{G}_t = \{\mathbf{V}_t, \mathbf{E}_t\, \mathbf{X}_t\ \}$ *is the graph at timestamp* $t \in [0, T]$*, where* $\mathbf{V}_t \in \mathbb{R}^{|\mathbf{V}_t|}$*,* $\mathbf{E}_t \in \mathbb{R}^{|\mathbf{V}_t| \times |\mathbf{V}_t|}$ *are the set of nodes and edges in* $\mathbf{G}_t$ *and* $\mathbf{X}_t \in \mathbb{R}^{|\mathbf{V}_t| \times L}$ *is the text feature of nodes in* $\mathbf{G}_t$ *where* L *is the maximum text length. Note that* $\mathbf{X}_t$ *stores the raw text content from each web domain and can be then processed into text embedding vectors for models to use.*

Definition 2 (Temporal Node Regression Task) *Let* $\mathbf{G}_t \in \mathbf{G}$ *be snapshot* t *of TAG* $\mathbf{G}$ *and* $\mathbf{V}_t$ *is the set of vertices in* $\mathbf{G}_t$*. Let* y *be the true labeling function over nodes in* $\mathbf{V}_t$*, where* $y : \mathbf{V}_t \to [0, 1]$*. The goal of the temporal node regression task is to learn a function* $f : \mathbf{V}_t \to [0, 1]$ *that approximates* y*:*

$$f(v) = y(v); \forall v \in \mathbf{V}_t$$

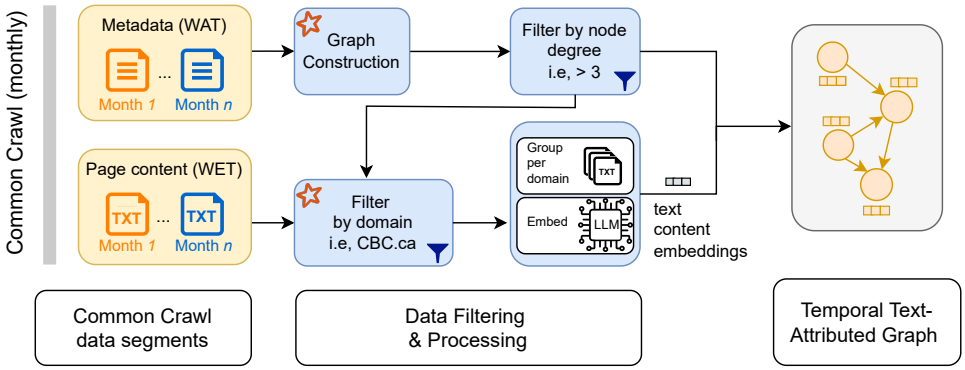


Figure 2: The temporal graph construction pipeline is implemented on a PySpark cluster to process Common Crawl domains collected across multiple months. WAT file segments are used to construct the domain-level web graph, with nodes of degree ≤ 3 excluded. Domain-level textual content is extracted from WET files, aggregated, and embedded with an off-the-shelf large language model to initialize node features. Edge timestamps are assigned based on the first day of the week of the crawl (usually, the first week of the month).

Note that across multiple timestamps, the label of a node might evolve as well.

Domain credibility assignment task. The domain credibility assignment task is the main temporal node regression task we support on CrediBench. The Credibility scores are assigned to the graph by matching domain names between the graph nodes and the entities of the Domain Quality Ratings (DQR) dataset [37]. The DQR dataset comprises approximately 11.5K web domains annotated with credibility scores collected from six different expert sources, including media organizations and independent professional fact-checkers. These sources apply different scoring schemes and evaluate domains across multiple dimensions (e.g., trust, reliability, bias, and transparency). Among these scores, we use the Media Bias/Fact Check (MBFC) score [2], which relies on human fact-checkers from the International Fact-Checking Network to assess media outlets on criteria such as factual accuracy and bias. In addition, we employ the first principal component (PC1), introduced by the DQR and derived from a principal component analysis (PCA) of the six scoring sources. The PC1 serves as an aggregate credibility measure that captures the common signal across the diverse sources, thereby mitigating variations between them. Both of these measures are in the range of $[0, 1]$ where a higher score indicates better credibility and less misinformation.

4 CrediBench

Figure 2 shows the detailed CrediBench pipeline for constructing our temporal text-attributed web graphs, where a one-month snapshot may contain upwards of a billion edges. The data provided by Common Crawl is in Web ARChive (WARC) files [1], the industry standard for web archiving. Metadata used to build the graph is found in Web ARChive Timestamp (WAT) files, containing HTTP response headers, links extracted from HTML pages and other metadata. Additionally, WARC Encapsulated Text (WET) files contain extracted plain text from web content.

Graph construction. The first step in graph construction is to download and decompress the raw data based on Common Crawl's decompression pipeline [16]. The downloaded files contains about 90,000 instances for one month, which are iteratively processed in batches of 300. After decompressing the files, the metadata and domain hyperlinks are extracted, resulting in a web graph where each node is a web domain, and each edge a hyperlink from one web domain's page to another. When aggregating over batches, deduplication is performed on the list of domains while subdomains and their parent domain are counted as separate nodes, i.e., `domain.com` and `domain.com.news`.

Domain filtering. Following the entire graph's decompression and construction, we attach credibility scores when available from [37], as well as timestamps at a monthly granularity (as provided in Common Crawl web graphs). We then filter domains in the graph to discard nodes with degree below a threshold which we set to 3. This processing step is motivated by the fact that lower-degree (especially isolated) nodes are less likely to be relevant during retrieval queries or generally in

information searches, and removing them lightens the computational load of handling the graph. A threshold 0 effectively means eliminating all isolated nodes, that of 1 eliminates also leaves, and so on. A threshold of 3, as we set it here, seeks a balance between this and the information loss it entails.

Text content extraction. Given the list of domains present in the graph, we extract text content for the web domains by loading their scraped content provided by Common Crawl WET file format, which stores the extracted plain body text of pages from the original HTML. A distributed Spark cluster is employed to process approximately 7.3 TB of WET files [18] in batches adjustable to the available compute resources. Each job extracts the text content, scraping timestamp, content languages, and corresponding webpage URL for the subsequent text classification task. The number of textual documents associated with each domain varies according to its update frequency; for instance, in the December 2024 snapshot, `github.com` contains approximately 20,000 documents, whereas other domains may contain only a single document. In the document grouping step, all documents belonging to a given domain alongside their timestamps are appended into a single document to facilitate subsequent embedding generation. To construct representative samples, we retain the three longest and three shortest documents per domain and merge them for embedding. For domains lacking textual content, we employ a multi-threaded online scraping pipeline in batches to extract text directly from the domain's home page.

Example December 2024 Snapshot. To explore the structural characteristics of credible websites, we utilize the CrediBench loading pipeline to yield a one-month snapshot from December 2024 [2]. Table 1 details the structure of the December 2024 raw and processed web graphs constructed through the CrediBench pipeline; where the processed one has only nodes with degree strictly higher than 3. We use the latter for experiments in this paper. Processing the graph as such keeps 90.21% of edges and 33.98% of nodes, thus slightly increasing the graph's density. The final processed file size of edges is 8.78 GB.

Table 1: Features of the December 2024 web graph.

Feature	Raw	Processed
$\|V\|$	132,547,562	45,041,648
$\|E\|$	1,124,576,420	1,014,523,552
Isolated nodes (deg = 0)	1,404,051	11,395
Leaves (deg = 1)	67,946,252	28,857
Edge density	1.28e-07	1.00e-06
Min. degree	0	0
Max. degree	14,900,588	14,719,077
Mean degree	16.97	45.05

In this paper, our analysis centers on a one-month snapshot, and these initial results already highlight the strength of such patterns in signaling credibility. Nevertheless, we anticipate that temporal dynamics will exert an equally significant influence, given the inherently evolving nature of web domains and online information propagation. Preliminary analyses of subgraphs from October and November 2024 further underscore this expectation, revealing clear signs of structural evolution. For instance, up to 40% of overlapping nodes experience an increase in out-degree from one month to the next. These observations motivate future research aimed at systematically examining the role of dynamic patterns in shaping credibility signals.

5 Experiments

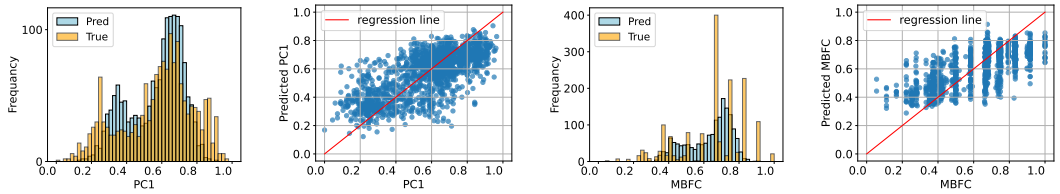
We experimented on the December 2024 snapshot to gain insights on signals that can infer the credibility of a domain. We aim to answer the following three research questions in our experiments:

- **RQ1:** *to what extent does the text content of a domain indicates its credibility score?*
- **RQ2:** *to what extent does the hyperlink structure of a web domain inform its credibility score?*
- **RQ3:** *to what extent can we benefit from combining the graph structure with the text content?*

Dataset: We use the domain quality rating dataset DQR [37] and its extracted textual content to train and test all models. We split the dataset into 60%, 20%, and 20% for train, validation, and test set splits. The regression targets on the DQR datasets are the PC1 and MBFC scores.

Compared Methods. To empirically examine all three research questions, we include a range of models to learn the text and graph modalities in our dataset, including the following:

[2]With Crawl ID `CC-MAIN-2024-51`.



(a) PC1 Score Distribution.

(b) PC1 regression scatter plot.

(c) MBFC Score Distribution.

(d) MBFC regression scatter plot.

Figure 3: Domain text provides the MLP with a clear signal for predicting the PC1 and MBFC scores.

- **Text-Based Models:** We employ the classic **TF-IDF** representation to embed each domain's textual content. The resulting embeddings are used as input to Multi-Layer Perceptron (MLP) decoder to predict the score.
- **LLM-based Models:** We also use various LLM models of different sizes to embed the text content into a latent vector, including: *EmbeddingGemma-300M* [21], *Qwen3 (0.6B and 8B)* embedding models [53], and *OpenAI Text Embedding 3 Large (OpenAI-TE3L)* [3]. Then, the text embeddings are fed to the MLP regressor for score prediction.
- **Graph-Based Models:** We utilize several Graph Neural Network (GNN) models: GCN[33], GraphSAGE[27], GAT[50], and GATv2 [10], to learn the hyperlink structure among domains and predict credibility scores.
- **Graph + Text Model:** We combine graph embeddings(from GAT) and text embeddings (from OpenAI-TE3L) to form joint feature representations. These combined embeddings are fed into an MLP regressor for credibility prediction. In the paper, refer to the *GAT+OpenAI-TE3L+MLP* model as the *graph+text* model.

We report regression results for the PC1 and MBFC scores in Table 2. We also include the simple mean baseline for comparison, which always predicts the mean score of all domains. To better understand models' performance, we report both the Mean Absolute Error (MAE) as well as the Maximum Absolute Error (Max AE) in both tables. The MAE reflects overall performance across many domains, while the Max AE reflects the worst possible deviation from the ground-truth score.

RQ1: Regression with text content. Here, we aim to answer if the extracted web text content contains beneficial signals for predicting domain credibility scores. This question is reflected by the performance of *text-based* models and *LLM-based* models as shown in Table 2. All models using learned text features outperform the mean baseline for both scores. This shows that the text content in CrediBench provides significant signals for credibility prediction when embedded via various methods (LLM or TF-IDF). Across both PC1 and MBFC scores, we observe that the LLM with the most parameters, i.e. OpenAI-TE3L, is the best performer in the text-only category. In addition, within the same Qwen3 family, more parameters also enables significantly better performance. We observe up to 0.014 MAE improvement from Qwen3-0.6B to Qwen3-8b on the PC1 score and up to 0.008 MAE improvement on the MBFC score. Figure 3 highlights a noticeable signal between the textual content of web domains and their assigned credibility scores. The Mean Absolute Error (MAE) for the test set PC1 and MBFC are 0.119 and 0.105, compared to 0.167 and 0.153 with the mean predictor. Therefore, incorporating the temporal evolution of domain content into CrediBench appears to be a promising direction. Figures 3a and 3c illustrate how the textual content facilitated the model in learning the distribution of credibility scores. Additionally, Figures 3b and 3d depict the deviations of predicted scores from the true scores, reflecting the achieved MAE performance.

RQ2: Regression with graph structure. To understand the impact of graph structure in modeling credibility scores, we train Graph Neural Networks (GNNs) on the node regression task on the December 2024 snapshot. For all GNN models, we use the Random Node Initialization [4] (RNI) as the node feature, normally distributed in $\mathcal{N}(0, 1)$. Our experiments include four common GNN architectures: two classical GNNs, GCN [33] and GraphSAGE (SAGE) [27], and two attention-based GNNs, Graph Attention Network (GAT) [50] and GATv2 [10]. In table 2, we report the average of 3 random seeds. We observe that with RNI initialization, GNNs consistently outperform the mean baseline. In particular, we see that the GAT model provides the lowest MAE in both the PC1 score as well as the MBFC score under graph based models. This indicates that the hyperlink graph structure provides useful signal for credibility prediction. Additionally, we also experimented with using

Table 2: Performance comparison (MAE; lower is better) of graph, text and LLM based models. Best **first** and second results are highlighted in bold.

	Score Type	PC1		MBFC	
	Metric	**MAE**	**Max(AE)**	**MAE**	**Max(AE)**
	Mean	0.167	0.546	0.153	0.561
Text-only	Text TF-IDF	0.130$\pm$0.001	**0.484**	0.117$\pm$0.001	**0.414**
	Gemma-300M + MLP	0.142$\pm$0.001	0.557	0.119$\pm$0.001	0.536
	Qwen3-0.6B + MLP	0.137$\pm$0.001	0.602	0.122$\pm$0.001	0.536
	Qwen3-8B + MLP	0.135$\pm$0.001	0.516	0.116$\pm$0.001	0.454
	OpenAI-TE3L + MLP	0.119$\pm$0.001	0.514	0.105$\pm$0.001	0.770
Graph-only	GCN	0.130$\pm$0.001	0.837	0.118$\pm$0.001	0.766
	GraphSAGE	0.152$\pm$0.002	0.852	0.135$\pm$0.004	0.892
	GAT	0.128$\pm$0.002	0.806	0.115$\pm$0.020	0.765
	GATv2	0.130$\pm$0.001	0.793	0.117$\pm$0.001	0.830
	graph + text	**0.117**$\pm$0.0001	0.523	**0.101**$\pm$0.0001	0.471

zero initialization instead of RNI as initial node features for GNNs, we observe a sharp decline of performance, we believe that in this snapshot, the increased expressiveness from RNI is beneficial for GNNs in learning the task, detailed results are reported in Appendix F. For further ablations, we report the effect of varying the number of neighbours and the number of hops in sampling Appendix G, we find that using more hops with 10-50 neighbours sampled achieves better performance. While maintaining one-hop, but with an overall equal number of neighbours produces worse performance.

RQ3: Combining graph and text. Since the dataset provides rich graph and text features, combining both may achieve the best performance. By concatenating the graph embeddings from GAT and text embeddings from OpenAI-TE3L LLM to form an input representation to a task MLP, the resulting graph+text model is trained to combine both modalities. Table 2 shows that the graph+text model achieves the best MAE. The graph+text MLP learns to extract complementary yet strong signals from both modalities. This is a promising first look into how to leverage the unique and novel features in CrediBench datasets and future work can further extend on this idea.

6 Conclusion

In this work, we introduced CrediBench, a large-scale web graph construction pipeline accompanied by a 1-month snapshot obtained through it, on which we explore the strength of different signals, including structural and text content, in learning credibility scores. Indeed, our empirical findings show that it is important to model both the text and graph structures of web domains in order to achieve better credibility score assignment. Our results strongly motivate further work to incorporate these signals, which in turn motivates the exploration of temporal aspects in dynamic evolution as greater leverage for predicting credibility scoring of web domains. In turn, such scores should facilitate the tracking and detection of misinformation at scale. By providing this pipeline and the resulting dataset, we aim to advance the field of misinformation detection, which is crucial in today's digital age where LLM-generated content is increasingly permeating digital ecosystems.

Acknowledgments and Disclosure of Funding

This research was supported by the Engineering and Physical Sciences Research Council (EPSRC) and the AI Security Institute (AISI) grant: Towards Trustworthy AI Agents for Information Veracity and the EPSRC Turing AI World-Leading Research Fellowship No. EP/X040062/1 and EPSRC AI Hub No. EP/Y028872/1. This research was also enabled in part by compute resources provided by Mila (mila.quebec) and Compute Canada.

References

[1] WARC (Web ARChive) File Format. Digital format description, Library of Congress, 2024. URL https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml. Last significant FDD update: April 29, 2024.

[2] Media Bias/Fact Check. 2023. Media Bias/Fact Check methodology (Internet Archive), 2025. URL https://web.archive.org/web/20230502031920/https://mediabiasfactcheck.com/methodology/. Last accessed: Sep 4, 2025.

[3] Text embedding 3 large, 2025. URL https://platform.openai.com/docs/models/text-embedding-3-large/. Last accessed: Oct 4, 2025.

[4] Ralph Abboud, İsmail İlkan Ceylan, Martin Grohe, and Thomas Lukasiewicz. The Surprising Power of Graph Neural Networks with Random Node Initialization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 2112–2118. ijcai.org, 2021. doi: 10.24963/IJCAI.2021/291. URL https://doi.org/10.24963/ijcai.2021/291.

[5] Gleb Bazhenov, Oleg Platonov, and Liudmila Prokhorenkova. GraphLand: Evaluating Graph Machine Learning Models on Diverse Industrial Data, 2025. URL https://arxiv.org/abs/2409.14500.

[6] Maya Bechler-Speicher, Ben Finkelshtein, Fabrizio Frasca, Luis Müller, Jan Tönshoff, Antoine Siraudin, Viktor Zaverkin, Michael M. Bronstein, Mathias Niepert, Bryan Perozzi, Mikhail Galkin, and Christopher Morris. Position: Graph Learning Will Lose Relevance Due To Poor Benchmarks. *CoRR*, abs/2502.14546, 2025. doi: 10.48550/ARXIV.2502.14546. URL https://doi.org/10.48550/arXiv.2502.14546.

[7] Sarah Blakeslee. The CRAAP Test. *LOEX Quarterly*, 2004.

[8] Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. DEFAME: Dynamic Evidence-based FAct-checking with Multimodal Experts. *CoRR*, abs/2412.10510, 2024. doi: 10.48550/ARXIV.2412.10510. URL https://doi.org/10.48550/arXiv.2412.10510.

[9] Sergey Brin and Lawrence Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Comput. Networks*, 30(1-7):107–117, 1998. doi: 10.1016/S0169-7552(98)00110-X. URL https://doi.org/10.1016/S0169-7552(98)00110-X.

[10] Shaked Brody, Uri Alon, and Eran Yahav. How Attentive are Graph Attention Networks? In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=F72ximsx7C1.

[11] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 675–684. ACM, 2011. doi: 10.1145/1963405.1963500. URL https://doi.org/10.1145/1963405.1963500.

[12] Mike Caulfield. SIFT (The Four Moves). *Hapgood*, 2019.

[13] Canyu Chen and Kai Shu. Can LLM-Generated Misinformation Be Detected? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=ccxD4mtkTU.

[14] Canyu Chen and Kai Shu. Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Mag.*, 45(3):354–368, 2024. doi: 10.1002/AAAI.12188. URL https://doi.org/10.1002/aaai.12188.

[15] Anshika Choudhary and Anuja Arora. Linguistic feature based learning model for fake news detection and classification. *Expert Syst. Appl.*, 169:114171, 2021. doi: 10.1016/J.ESWA.2020.114171. URL https://doi.org/10.1016/j.eswa.2020.114171.

[16] Common Crawl. Common Crawl PySpark Repository. URL https://github.com/commoncrawl/cc-pyspark.

[17] Common Crawl. Common Crawl Web Graphs, 2025. URL https://commoncrawl.org/web-graphs.

[18] Common Crawl. December 2024 Crawl Archive Now Available, 2025. URL https://commoncrawl.org/blog/december-2024-crawl-archive-now-available.

[19] Mark Elsner, Grace Atkinson, and Saadia Zahidi. Global Risks Report 2025, Jan. 2025. URL https://www.weforum.org/publications/global-risks-report-2025/. 20th edition of the Global Risks Report.

[20] Kenneth C. Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzeminski, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çagatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafal Poswiata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Suppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal A. Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James V. Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. MMTEB: Massive Multilingual Text Embedding Benchmark. *CoRR*, abs/2502.13595, 2025. doi: 10.48550/ARXIV.2502.13595. URL https://doi.org/10.48550/arXiv.2502.13595.

[21] Henrique Schechter Vera et al. Embeddinggemma: Powerful and lightweight text representations, 2025. URL https://arxiv.org/abs/2509.20354.

[22] Gunther Eysenbauch. Infodemiology: the epidemiology of (mis)information. 113(9). URL https://www.amjmed.com/article/S0002-9343(02)01473-0/fulltext.

[23] Matthias Fey and Jan Eric Lenssen. Fast Graph Representation Learning with PyTorch Geometric. *CoRR*, abs/1903.02428, 2019. URL http://arxiv.org/abs/1903.02428.

[24] Matthias Fey, Jinu Sunil, Akihiro Nitta, Rishi Puri, Manan Shah, Blaz Stojanovic, Ramona Bendias, Alexandria Barghi, Vid Kocijan, Zecheng Zhang, Xinwei He, Jan Eric Lenssen, and Jure Leskovec. PyG 2.0: Scalable Learning on Real World Graphs. *CoRR*, abs/2507.16991, 2025. doi: 10.48550/ARXIV.2507.16991. URL https://doi.org/10.48550/arXiv.2507.16991.

[25] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. TweetCred: Real-Time Credibility Assessment of Content on Twitter. In *Social Informatics - 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings*, volume 8851 of *Lecture Notes in Computer Science*, pages 228–243. Springer, 2014. doi: 10.1007/978-3-319-13734-6_16. URL https://doi.org/10.1007/978-3-319-13734-6_16.

[26] Suhaib Kh Hamed, Mohd Juzaiddin Ab Aziz, and Mohd Ridzwan Yaakub. A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. *Heliyon*, 2023.

[27] William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 1024–1034, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html.

[28] Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. Compare to the Knowledge: Graph Neural Fake News Detection with External Knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.62. URL https://aclanthology.org/2021.acl-long.62/.

[29] Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. Disinformation Detection: An Evolving Challenge in the Age of LLMs. URL https://www.semanticscholar.org/paper/Disinformation-Detection%3A-An-Evolving-Challenge-in-Jiang-Tan/6b1c431db1f7d10f0a55d51786d55ad6b6921730.

[30] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multim. Tools Appl.*, 80(8): 11765–11788, 2021. doi: 10.1007/S11042-020-10183-2. URL https://doi.org/10.1007/s11042-020-10183-2.

[31] Seyed Mehran Kazemi, Rishab Goel, Kshitij Jain, Ivan Kobyzev, Akshay Sethi, Peter Forsyth, and Pascal Poupart. Representation learning for dynamic graphs: A survey. *J. Mach. Learn. Res.*, 21:70:1–70:73, 2020. URL https://jmlr.org/papers/v21/19-447.html.

[32] Hamid Keshavarz. Evaluating credibility of social media information: current challenges, research directions and practical criteria. URL https://www.emerald.com/insight/content/doi/10.1108/idd-03-2020-0033/full/html.

[33] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

[34] Rina Kumari and Asif Ekbal. AMFB: Attention based multimodal Factorized Bilinear Pooling for multimodal Fake News Detection. *Expert Syst. Appl.*, 184:115412, 2021. doi: 10.1016/J.ESWA.2021.115412. URL https://doi.org/10.1016/j.eswa.2021.115412.

[35] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham M. Kakade, Prateek Jain, and Ali Farhadi. Matryoshka Representation Learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/c32319f4868da7613d78af9993100e42-Abstract-Conference.html.

[36] Batool Lakzaei, Mostafa Haghir Chehreghani, and Alireza Bagheri. A Decision-Based Heterogenous Graph Attention Network for Multi-Class Fake News Detection. *CoRR*, abs/2501.03290, 2025. doi: 10.48550/ARXIV.2501.03290. URL https://doi.org/10.48550/arXiv.2501.03290.

[37] Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David G Rand, and Gordon Pennycook. High level of correspondence across different news domain quality rating sets. *PNAS Nexus*, 2(9):pgad286, 09 2023. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgad286. URL https://doi.org/10.1093/pnasnexus/pgad286.

[38] Priyanka Meel and Dinesh Kumar Vishwakarma. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Syst. Appl.*, 153:112986, 2020. doi: 10.1016/J.ESWA.2019.112986. URL https://doi.org/10.1016/j.eswa.2019.112986.

[39] Hejamadi Rama Moorthy, N. J. Avinash, Krishnaraj N. S. Rao, K. R. Raghunandan, Radhakrishna Dodmane, Jeremy Joseph Blum, and Lubna A Gabralla. Dual stream graph augmented transformer model integrating BERT and GNNs for context aware fake news detection. 15. URL https://www.nature.com/articles/s41598-025-05586-w.

[40] Mohammad Vatani Nezafat and Saeed Samet. Fake News Detection with Retrieval Augmented Generative Artificial Intelligence. In *2nd International Conference on Foundation and Large Language Models, FLLM 2024, Dubai, United Arab Emirates, November 26-29, 2024*, pages 160–167. IEEE, 2024. doi: 10.1109/FLLM63129.2024.10852474. URL https://doi.org/10.1109/FLLM63129.2024.10852474.

[41] John O'Donovan, Byungkyu Kang, Greg Meyer, Tobias Höllerer, and Sibel Adali. Credibility in Context: An Analysis of Feature Distributions in Twitter. In *2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Confernece on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012*, pages 293–301. IEEE Computer Society, 2012. doi: 10.1109/SOCIALCOM-PASSAT.2012.128. URL https://doi.org/10.1109/SocialCom-PASSAT.2012.128.

[42] Alexandra Olteanu, Stanislav Peshterliev, Xin Liu, and Karl Aberer. Web Credibility: Features Exploration and Credibility Prediction. In *Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings*, volume 7814 of *Lecture Notes in Computer Science*, pages 557–568. Springer, 2013. doi: 10.1007/978-3-642-36973-5_47. URL https://doi.org/10.1007/978-3-642-36973-5_47.

[43] William Scott Paka, Rachit Bansal, Abhay Kaushik, Shubhashis Sengupta, and Tanmoy Chakraborty. Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection. *Appl. Soft Comput.*, 107:107393, 2021. doi: 10.1016/J.ASOC.2021.107393. URL https://doi.org/10.1016/j.asoc.2021.107393.

[44] Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. Towards Reliable Misinformation Mitigation: Generalization, Uncertainty, and GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6399–6429. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.395. URL https://doi.org/10.18653/v1/2023.emnlp-main.395.

[45] Yuxiang Ren, Bo Wang, Jiawei Zhang, and Yi Chang. Adversarial Active Learning based Heterogeneous Graph Neural Network for Fake News Detection. In *20th IEEE International Conference on Data Mining, ICDM 2020, Sorrento, Italy, November 17-20, 2020*, pages 452–461. IEEE, 2020. doi: 10.1109/ICDM50108.2020.00054. URL https://doi.org/10.1109/ICDM50108.2020.00054.

[46] Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI: A Hybrid Deep Model for Fake News Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 797–806. ACM, 2017. doi: 10.1145/3132847.3132877. URL https://doi.org/10.1145/3132847.3132877.

[47] Abhijit Suprem, Sanjyot Vaidya, and Calton Pu. Exploring Generalizability of Fine-Tuned Models for Fake News Detection. In *8th IEEE International Conference on Collaboration and Internet Computing, CIC 2022, Atlanta, GA, USA, December 14-16, 2022*, pages 82–88. IEEE, 2022. doi: 10.1109/CIC56439.2022.00022. URL https://doi.org/10.1109/CIC56439.2022.00022.

[48] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. Some Like it Hoax: Automated Fake News Detection in Social Networks. *CoRR*, abs/1704.07506, 2017. URL http://arxiv.org/abs/1704.07506.

[49] Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera, Mayank Goel, Zachary Yang, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. Web Retrieval Agents for Evidence-Based Misinformation Detection. *CoRR*, abs/2409.00009, 2024. doi: 10.48550/ARXIV.2409.00009. URL https://doi.org/10.48550/arXiv.2409.00009.

[50] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference*

Track Proceedings. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

[51] William Yang Wang. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 422–426. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-2067. URL https://doi.org/10.18653/v1/P17-2067.

[52] Xin Xia, Xiaohu Yang, Chao Wu, Shanping Li, and Linfeng Bao. Information Credibility on Twitter in Emergency Situation. In *Intelligence and Security Informatics - Pacific Asia Workshop, PAISI 2012, Kuala Lumpur, Malaysia, May 29, 2012. Proceedings*, volume 7299 of *Lecture Notes in Computer Science*, pages 45–59. Springer, 2012. doi: 10.1007/978-3-642-30428-6_4. URL https://doi.org/10.1007/978-3-642-30428-6_4.

[53] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

[54] Jiawei Zhang, Bowen Dong, and Philip S. Yu. FakeDetector: Effective Fake News Detection with Deep Diffusive Neural Network. In *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020*, pages 1826–1829. IEEE, 2020. doi: 10.1109/ICDE48307.2020.00180. URL https://doi.org/10.1109/ICDE48307.2020.00180.

[55] Mengfan Zhao, Yutao Zhang, and Guozheng Rao. Fake news detection based on dual-channel graph convolutional attention network. *J. Supercomput.*, 80(9):13250–13271, 2024. doi: 10.1007/S11227-024-05953-W. URL https://doi.org/10.1007/s11227-024-05953-w.

A Limitations

Relying on Common Crawl web graphs as its data source, the CrediBench pipeline may inherit some of its features, which we cannot verify. Coverage may be biased, for example, capturing domains unequally, or noisy text content, all of which may affect the quality of our data and limit the strength of insights drawn from it. Similarly, with our small labelled set relative to all the domains, we are limited in the credibility we can effectively infer. Moreover, we recognize that, despite our best efforts, considering the large scale of the data, running the CrediBench can be expensive in its computational requirements. With this consideration, we will release our curated December 2024 snapshot after the review period for ease of access. Finally, our current experiments are based on the December 2024 snapshot only, which already contains billions of edges. We are actively running CrediBench to extract additional snapshots and will share publicly for research purposes once they are available.

B Broader Impact

Positive impact in graph learning. Misinformation is becoming more and more prevalent online. Therefore, it is important to design ML methods to detect misinformation domains. To this end, CrediBench constructs a dataset which benefits the community by introducing better baseline datasets for misinformation detection. Additionally, this dataset provides a large-scale real-world dataset for graph ML, which is beneficial as a benchmark for the field.

Negative societal impact. The nature of credibility score assignment can bring potential negative impact, especially if a learned model incorrectly assigned low or high credibility score. In addition,

credibility and misinformation can be subjective depending on the individual; thus it is important to be aware of such impact when deploying a ML system for misinformation detection. Lastly, due to the nature of our labels on the web domains, they are biased towards news websites, thus focusing more on a certain type of domains.

C Compute

For the data downloading, decompressing and graph construction, constructing one monthly snapshot takes around 600 hours for all batches on a single NVIDIA RTX8000 with 265GB of RAM and 4 CPU cores. Downstream degree-based filtering and target label generation for supervised learning consume an additional 7 hours per monthly snapshot of upwards of a billion edges, with the same computational setup. To enable this, this step offloads sorting and joins to disk-backed external processed and temporary files to avoid costly in-python accumulators. We also use memory-mapped degree vectors to avoid loading large structures into RAM, and performing streaming, sequential passes that minimize in-memory state. The result pipeline processes massive graphs efficiently while remaining resource-bounded and reproducible.

Extracting a monthly snapshot of textual content requires approximately 100 hours on a single high-performance virtual machine equipped with 256 GB of RAM and 64 vCPU cores. The content embeddings for the DQR dataset domains were generated separately on a V100D-16C vGPU with 16 GB of VRAM, requiring about 2 hours. To facilitate GNN on our large-scale graph, we use a neighbour loader sampler (sampling multi-hop neighbours) to sample adequate training batches of size 50 at 3-hops ($[50, 50, 50]$). The experiments were run using a 80GB memory NVIDIA A100 Tensor Core GPU.

D Text content

Text embedding generation. The content of each domain is transformed into an embedding using the Qwen3-0.6B multilingual embedding model [53], which produces 1024-dimensional representations per domain that could be downsized into smaller dimensions (i.e., 128) as it supports Matryoshka Representation Learning (MRL) [35]. Notably, Qwen3-0.6B ranks fourth on the MTEB leaderboard while being the smallest model among the top-performing models [20]. Figure 4 shows the relative sizes of the Common Crawl December-2024 webgraph that consists of 132.5M nodes, and the processed subgraph that contains 45M nodes, while the DQR labeled dataset contains only 11.5K nodes.

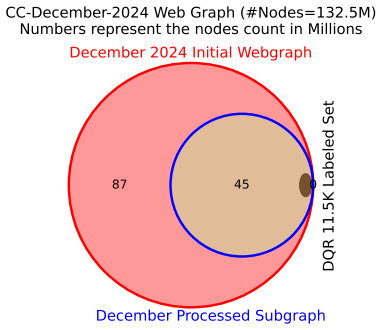


Figure 4: The relative sizes of CC-December 2024 webgraph, the processed subgraph, and the DQR.

Domain content examples with their PC1 and MBFC scores. The following examples present text content extracted from CC-WET files for various DQR domains, along with their corresponding credibility scores. The higher the score, the more credible the domain.

Domain Name: ncdc.noaa.gov	**PC1 Score: 0.90**	**MBFC Score: 0.84**

News | National Centers for Environmental Information (NCEI)
Skip to main content
An official website of the United States government
Here's how you know
Federal government websites often end in .gov or .mil. Before sharing sensitive information, make sure you're on a federal government site.
There's a lot going on at NCEI. Discover more about us and Earth's climate, oceans, coasts, and geophysics in these featured news stories.
October 4, 2024
Helene Devastates Southeast, Impacts NOAA NCEI Headquarters
NCEI headquarters in Asheville, NC, has been severely impacted by Hurricane Helene. All of our employees and staff have been accounted for with all data holdings–including paper and film records–safe.
Read More ....

Domain Name: ctvnews.ca	**PC1 Score: 0.92**	**MBFC Score: 0.91**

How many people in Canada have died from COVID? | CTV News
FOLLOW ON CORONAVIRUS UPDATES Complete coverage at CTVNews.ca/Coronavirus
COVID-19 NEWSLETTER
Text:
July 19, 2022 update: This tracker is no longer being maintained but it will continue to be available as an archive. As the collection and dissemination of provincial COVID-19 data has evolved over time, we've made available a new version of the tracker here.
Jan. 14, 2022 update: The CTV News coronavirus tracker is now highlighting hospitalizations and ICU admissions across Canada, with an interactive map that includes provincial breakdowns. Our original tracker which has been keeping count of cases since March 2020 hasn't gone anywhere, it's right below the hospitalizations.
Numbers on map reflect total hospitalizations including ICU admissions. Line graph below breaks down hospitalization and ICU admissions for each province and territory
Across Canada-Cases-Total
...

Domain Name: nbcsports.com **PC1 Score: 0.76** **MBFC Score: 0.79**

Login | NBC Sports - NBC Sports
Skip navigation
Search Query Submit Search
MLB NFL NBA NHL NASCAR
....
Top News
NASCAR Cup starting lineup at Talladega: Michael McDowell wins pole
Dustin Long,
Tyrrell Hatton ties Old Course record to lead; Nicolas Colsaerts one back after albatross
Associated Press,
WATCH: Dog steals Gareth Bale's ball on green at Dunhill Links
Golf Channel Staff,
Top Clips
Neville: 'The pressure is enormous' on ten Hag
Extended HLs: Everton v. Newcastle Matchweek 7
Talladega brings size, speed, tradition to Alabama
...

Domain Name: foxnews.com **PC1 Score: 0.53** **MBFC Score: 0.58**

Tego Calderón Rocks Washington Heights | Fox News
Go Back Fox News Move Back ADVERTISEMENT Skip
Published December 9, 2016 - 6 Images
Tego Calderón Rocks Washington Heights
Tego Calderón took the stage in New York City's Highbridge Park and gave a free concert for the Washington Heights community.
Start Slideshow
read more Facebook Twitter Email Link

Domain Name: climate.news **PC1 Score: 0.07** **MBFC Score: 0.11**

Climate News | Climate News & Climate Studies
sustainable living
01/08/2018 / By Zoey Sky
A new kind of magic mushroom: New sustainable material made of mushrooms can provide housing, food security, water filtration
While some people are worried that we might one day run out of building materials, a group of experts has revealed that they have successfully created wood-like blocks out of mushrooms. Architect Chris Maurer and his collaborators at Redhouse Architecture in Cleveland, Ohio, aim for the creation of whole communities using mushroom "wood" and its [. . .]
« Return Home 1 of 1 Popular Posts
Climate.News is a fact-based public education website published by Climate News Features, LLC. All content copyright © 2018 by Climate News Features, LLC.
Contact Us with Tips or Corrections
All trademarks, registered trademarks and servicemarks mentioned on this site are the property of their respective owners.

Table 3: Performance comparison (MAE) of different graph neural network architectures (zero initializations) and MLP on **pc1-score** and **MBFC-score**. Qwen3-0.6B[53] text embeddings are used as MLP input features. Best results are highlighted in bold.

Method	PC1 (MAE)	MBFC (MAE)
Mean	0.167	0.153
Text only (MLP + Qwen3)	**0.137** ± **0.001**	**0.122** ± **0.001**
GCN (zero)	0.186 ± 0.030	0.254 ± 0.004
GraphSAGE (zero)	0.189 ± 0.028	0.252 ± 0.031
GAT (zero)	0.296 ± 0.051	0.153 ± 0.001
GATv2 (zero)	0.321 ± 0.040	0.302 ± 0.020

Domain Name: naturalnews.com **PC1 Score: 0** **MBFC Score: 0.11**

Preventing liver damage with the henna plant – NaturalNews.com
Home Brighteon Prep with Mike Interviews Audio Books
Download Our App About Us FAQs Search
Sections-Follow Us-Podcast-Store Subscribe-Home Politics Culture Health - Medicine Finance - Economy Prepping - Survival Science Technology-Popular Articles-Today Week
See More Popular Articles
Health Ranger Report
2:42:03 Brighteon Broadcast News, Oct 9, 2024 Hurricanes, fires and floods all designed to DESTROY cities, then REBUILD as open-air PRISON CAMPS
18:14 Unconfirmed reports: Israel preparing for pre-emptive NUCLEAR strike on Iran
26:26 FEMA setting a TRAP for Floridians to try to provoke a REVOLT
....

E Experimental Setup

The DQR dataset splitting. The DQR domains are stratified according to the target scores (i.e., PC1 and MBFC) and partitioned into training, validation, and test sets comprising 60%, 20%, and 20% of the data, respectively

MLP Experiment. A scikit-learn MLP regressor with two hidden layers (the first of dimension 128, the second 64) with a ReLU activation function, an Adam optimizer, and a learning rate of 0.001 is applied. The models were trained for 15 epochs with a maximum number of iterations set to 200.

GNN Experiment. We used different GNN architectures, such as residual connections, which performed well in node-regression tasks [5]. Each GNN has 3 layers, a dropout of 0.1, and a learning rate of 0.001 on the Adam optimizer. Due to the scale of our dataset, we employed the use of PyG's NeighborLoader [23, 24], a neighbor sampling method introduced in [27] to mini-batch our data. Each GNN runs for 100 epochs over 3 trials.

F Comparing RNI initialization with zero-vector embeddings

In Table 3, we report the performance of GNNs with zero initialization. As it highlights, this set-up does not outperform the mean baseline, as opposed to using RNI initializations. We also note that with this set-up, the standard deviation is generally increased.

G Ablation Studies on GNN experiments

In this section, we study the effect of the number of neighbors on the performance of GNNs for the PC1 score prediction. We report Out of Memory as OOM. The first observation is that generally as more number of neighbors are sampled, the GNN MAE performance on the PC1 score has improved. Similarly, in most cases, increasing the number of hops also improves the MAE score especially in

Table 4: Ablation study of GAT for PC1 score regression with varying numbers of neighbours sampled and hops. Large-scale one-hop settings are shown separately. Error is recorded in MAE.

Num Neighbors	**PC1 (MAE)**		
	1-hop	**2-hop**	**3-hop**
10,000	0.144 ± 0.001	N/A	N/A
125,000	0.145 ± 0.001	N/A	N/A
5	0.147 ± 0.001	0.148 ± 0.001	0.150 ± 0.002
10	0.141 ± 0.004	0.143 ± 0.003	0.138 ± 0.002
30	0.138 ± 0.003	0.137 ± 0.001	**0.129 ± 0.002**
50	0.142 ± 0.002	0.135 ± 0.001	OOM

the case of sampling 30 neighbors for each of three hops. This shows that multi-hop information and neighborhood information in the hyperlink graph helps with PC1 score prediction. We also conducted an experiment where a large number of 1 hop neighbors are sampled. In this case, we see that even with over 125k 1 hop neighbors, 3 hop information remains the best performer.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Main contributions and claims made are argued through experimentation or by providing data/code in the final submission.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Discussions on limitations can be found in the appendix section A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Models and their architecture used in the experiments are described in the results section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: the entire code used in our pipeline, graph construction, and experimentation is available here. The data is available in this folder.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: This information is reported with our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For all results, we report standard deviation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Detailed compute resources are documented in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: No potential harms.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Discussions on positive and negative societal impacts can be found in the appendix section B.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Poses no such risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Common Crawl (under open license) is credited and cited as appropriate.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code is documented in its repository's `README` file. The data will have a dataset card properly documenting it, upon release (that is, upon final submission).

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: N/A.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: N/A.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs were used our methodology to generate text embeddings for the web pages' content, as detailed in appendix D.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.