Infox-QC: A Quebec-Focused French Corpus for Misinformation Detection and AI Robustness Assessment

Moetaz Doghmane[1,3], **Hazem Amamou**[1,3], **Thiziri Sefsaf**[3], **Alan Davoust**[2,3], **Anderson R. Avila**[1,3]
[1]Institut national de la recherche scientifique (INRS-EMT), Montréal, Quebec, Canada
[2]Université du Québec en Outaouais, Gatineau, Quebec, Canada
[3]INRS-UQO Mixed Research Unit on Cybersecurity, Gatineau, Quebec, Canada
moetaz.doghmane@inrs.ca, hazem.amamou@inrs.ca, thiziri.sefsaf@gmail.com
alan.davoust@uqo.ca, anderson.avila@inrs.ca

Abstract

The pervasive spread of online misinformation, often through social media and political campaigns, makes detecting false claims a crucial task for mitigating societal risks. While the vast majority of fake news datasets are developed in English, a critical gap remains for low-resource languages, such as French. To address this, we introduce Infox-QC, a novel French-language corpus focused on misinformation relevant to the Quebec region. Beyond containing real true and fake news, Infox-QC includes two unique subsets of AI-generated fake news: one created by prompting an AI to paraphrase existing fake news, and a second generated by prompting an AI to fabricate fake news from real true reports. This innovative approach allows us to verify the robustness of detection systems against fabricated content, which modern LLMs can generate with convincing efficacy. We establish comprehensive baselines using traditional machine learning methods, BERT-based models, and Large Language Models, both with and without Retrieval-Augmented Generation (RAG). Our results demonstrate that RAG-augmented LLMs offer the strongest contextual understanding, while traditional models provide valuable interpretable baselines. We further provide an exploratory human–LLM thematic agreement analysis to assess annotation consistency. The Infox-QC resource fills a critical void in French-language NLP research, supporting future efforts to explore the regional and cultural dimensions of misinformation through cross-linguistic comparison.

Keywords: misinformation detection, French dataset, Quebec, machine learning, transformers, RAG

1. Introduction

Nowadays, misleading information comprises text, images, audio, video, or their combination. Such content is referred to as “misinformation” when there is no intention to harm, or "disinformation" if it represents the arbitrary intent to deceive people’s beliefs (Weatherall and O’Connor, 2024). While misleading campaigns to influence people’s decision dates back to centuries ago in Roman times (Posetti and Matthews, 2018), it was just recently with the advances in communication and ubiquity of social media that the rapid spread, at scale, of fake news became possible, posing urgent threats to democracies, especially if weaponized to sway public behavior and votes.

Fabricated information—such as fake news—designed to mimic legitimate news content and appear credible to readers is a key concern. Following Lazer et al. (2018), we define fake news as content produced without regard for accuracy and outside journalistic standards and organizational processes. It typically imitates the form of news (headlines, bylines, imagery) to mislead and erode trust, spreading rapidly through social platforms. Although the phenomenon of fake news has been studied for many years, the issue is far from being solved and new challenges raises daily (Plikynas et al., 2025), specially with the rapid progress of generative AI and widespread adoption of Large Language Models (LLMs). GPT-4 (used in ChatGPT) (Achiam et al., 2023), LLaMA (Touvron et al., 2023), and DeepSeek-R1 (DeepSeek-AI, 2025), for instance, can ultimately enable anyone to create highly convincing textual content. This poses a real threat to models trained solely on human-written fake news datasets.

We hypothesize that human- and LLM-generated fake news exhibit distinct linguistic patterns. Consequently, models trained only on human-written content may struggle to detect LLM-generated misinformation due to distribution shift. The magnitude of this gap remains unclear, especially in lower-resource languages such as French. To investigate this, we compare the performance of several classifiers on detecting original vs LLM-generated fake news. Our study delineates the challenges and limitations of multiple detectors—traditional machine learning, deep learning, and large language models, including RAG-based approaches. We focus on a corpus of fake news about Quebec that circulated between 2003 and 2025. Because misinformation narratives are shaped by local political and media contexts, focusing on Quebec allows us to study these dynamics within its sociocultural environment. Variants of French also differ across

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6979–6989
11–16 May 2026.

regions, making regionally grounded datasets an important complement to broader multilingual corpora. Thus, we provide a detailed qualitative analysis of the main content types and topics present in the collected data. Thus, our contributions are threefold:

- A French-language corpus of misinformation from Quebec, designed for fake news detection.
- The impact of AI-generated fake news on models trained with human-written texts.
- A comparative analysis of Machine Learning, LLM, and Retrieval-Augmented Generation (RAG) models for misinformation classification.

2. Related Work

Misinformation detection has become an increasingly active area of research. Although several benchmarks have been developed, most existing resources remain heavily English-centric, overlooking the cultural and linguistic nuances required to study misinformation across diverse contexts. This linguistic imbalance underscores the urgent need for high-quality datasets in low-resource languages such as French, where cultural, regional, and rhetorical differences play a crucial role in shaping misinformation narratives.

The LIAR dataset (Wang, 2017), one of the earliest benchmarks in the field, contains 12.8k political statements drawn from U.S. fact-checking websites. Despite its widespread adoption, its scope is narrow: it is almost entirely limited to political discourse, focuses solely on six levels of truthfulness without providing richer metadata such as audience type or dissemination platform, and is restricted to English, thereby reflecting U.S.-specific linguistic and cultural assumptions. Similarly, FakeNewsNet (Shu et al., 2020) is among the most comprehensive English misinformation corpora, with approximately 23,000 articles linked to social context information such as user engagement and propagation patterns. However, it also suffers from a strong U.S.-centric and English-only bias. Its thematic scope largely mirrors the American news cycle, and its annotations remain focused on article-level labels rather than human-centered dimensions such as audience targeting, impact estimation, or cross-platform distinctions.

While the Obsinfox dataset (Icard et al., 2024) represents a valuable contribution to French-language misinformation research, offering around 100 annotated articles with detailed textual labels (e.g., exaggeration, insinuation, framing), its limited scale, narrow temporal coverage, and exclusive focus on textual content constrain its utility for large-scale machine learning or broader sociological analyses.

3. The Infox-QC Dataset

Infox-QC is a French-language dataset about misinformation circulating in Quebec. It aims to provide a new resource to the misinformation and disinformation research community, featuring false information in the French language and specific to the region of Quebec in Canada. Besides representing region- and language-specific content, the dataset also aims to reflect the new reality of AI-generated content. In the next subsections, we provide details about the data collection procedure and the dataset characteristics.

3.1. Data Collection

The false information in the Infox-QC dataset was collected manually by the research team, from online sources. False news articles were retrieved using fact-checking reports and diverse social media channels, specialized in capturing the evolution of misinformation during events such as the COVID-19 pandemic and recent elections. These sources include fact-checking organizations and media outlets such as *Radio-Canada – Les Décrypteurs*, *AFP Factuel*, *Le Devoir*, and *Agence Science-Presse*, along with official public data sources (*Gouvernement du Québec*, *INSPQ*, *Santé Canada*). Additionally, social media platforms such as Facebook and X (Twitter) were monitored to identify and cross-verify circulating misinformation, among other comparable national and local sources. The process implied in gathering texts, eliminating redundancies, refining content to focus only on false claims, and thematic categorization (health, immigration, politics, climate). Iterative refinements enhanced textual data through pre-processing, ensuring alignment with thematic categories and verifying content accuracy, thereby establishing a robust basis for analyzing dissemination patterns and temporal trends.

Eliminating redundancies: During data cleaning, cosine similarity based on TF–IDF was used to identify items conveying the same false claim, with only minor wording differences. For similarity scores close to 1, only one text remained and the other were removed. Thus, we assured that the false claims were not repeated.

Refining content: After removing redundant items, we double-checked if the item conveyed a valid misinformation claims. Only items representing false information were considered valid and were kept as part of the data.

Iterative refinements: This means that the pre-processing and cleaning steps were performed in

Field	Description
id	A unique identifier assigned to each entry.
date	The identification date, standardized for temporal tracking.
content	A description of the statement, refined to isolate content for analysis.
type	The type of information (reportage, declaration, policy announcement, etc.).
theme	The thematic focus, which includes health, economy, environment, security, politics, etc.
objective	The possible intent (inform, mislead, manipulate, influence, etc.).
target audience	The intended audience, e.g., General Public, Youth, or specific ideological groups.
platform	The dissemination medium, covering platforms like social media and messaging services.
region	The geographic focus, primarily Quebec (75%), but also Canada and International items circulated in Quebec.
format	The content type, such as Text, Video, Image, or Deepfake.
truthfulness	The veracity level, indicating the extent of falsehood (e.g., TRUE or FALSE for binary classification).
source	The fact-checking authority, for example, established news outlets or agencies.
link	A reference link to the verification source.
impact	The estimated impact, rated as Low, Medium, High, or Very High.
language	The language, primarily French, with some bilingual entries.

Table 1: Metadata information categorizing each news article present in Infox-QC.

several iterations. Each iteration improved data consistency and thematic alignment.

To collect true statements, we adopted a multi-source strategy that blended automated tools and careful manual review. Automated retrieval relied on NewsAPI, MediaStack, and GNews API, while manual curation focused on trustworthy Quebec and Canadian media outlets. Credibility was assessed through journalistic transparency, editorial independence, and alignment with recognized fact-checking practices. Among the main sources consulted were La Presse, Le Devoir, Le Soleil, Journal de Montréal, Radio-Canada, and Global News. We also incorporated verified information from official public institutions such as the Gouvernement du Québec, INSPQ, Santé Québec, and Statistique Canada.

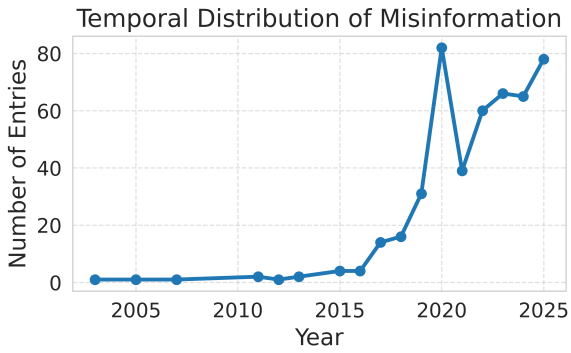


Figure 1: Temporal distribution of misinformation circulated in Quebec between 2003 and 2025.

3.2. Dataset Description

For each news article, Infox-QC provides the metadata listed in Table 1. As can be seen, it contains useful information for performing qualitative data analysis about misinformation circulated in Quebec. Besides binary labels necessary to training fake news detectors, such metadata contains contextual attributes such as the date of identification, type of content (e.g., reportage, declaration), format (e.g., Text, Video, Deepfake), and platform of diffusion. Crucially, the dataset provides essential sociological and geographical context by tagging the primary theme (e.g., health, politics), the article's objective (e.g., mislead, influence), the target audience (e.g., Youth, specific ideological groups), and its focus on the region of Quebec, thus enabling researchers to study the sociopolitical dynamics of French-language misinformation.

The dataset comprises 1,000 samples, evenly split between true and false news: 500 true news and 500 of false news. While modest in size compared to large English misinformation corpora, this scale reflects the careful manual verification and Québec-specific filtering applied during collection. The balanced design minimizes class bias and supports reliable benchmarking across model families. Furthermore, the inclusion of two AI-generated fake news subsets increases linguistic diversity and enables controlled robustness evaluation under distribution shift. Statements contain approximately 150 characters, consistent with the concise textual style of news headlines, political declarations, or social media posts. Figure 1 shows the temporal distribution of entries of the dataset, spanning the period from 2003 to 2025. Entries remained minimal during the 2003–2015 period but showed a sharp, event-driven escalation thereafter, as the concept of *fake news* gained prominence in the media landscape. The volume surged to its highest peak in 2020, coinciding with the COVID-19 pandemic, and has remained significantly elevated and

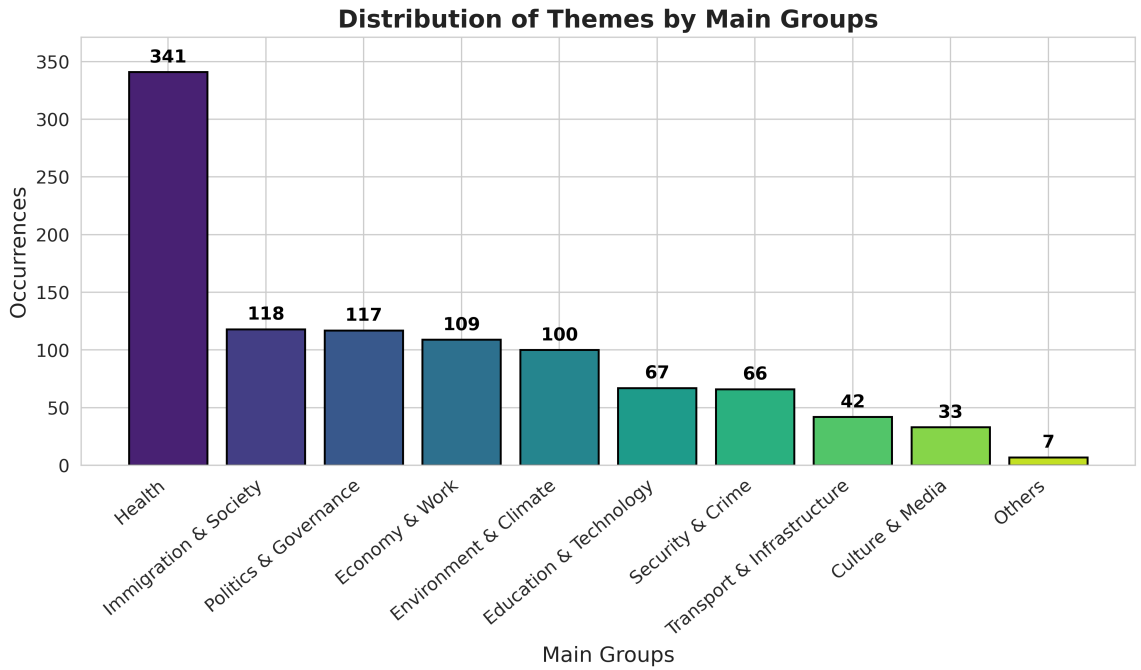


Figure 2: Distribution of thematic categories in the dataset.

LLM annotator	Accuracy	Cohen's κ	Macro-F1
GPT-4o-mini	0.796	0.754	0.688
Claude	0.767	0.721	0.671

Table 2: Agreement between human theme-group labels and LLM-assigned theme groups based solely on the textual claim field (1,000 samples).

fluctuating through 2025.

3.3. Theme and Topic Distribution

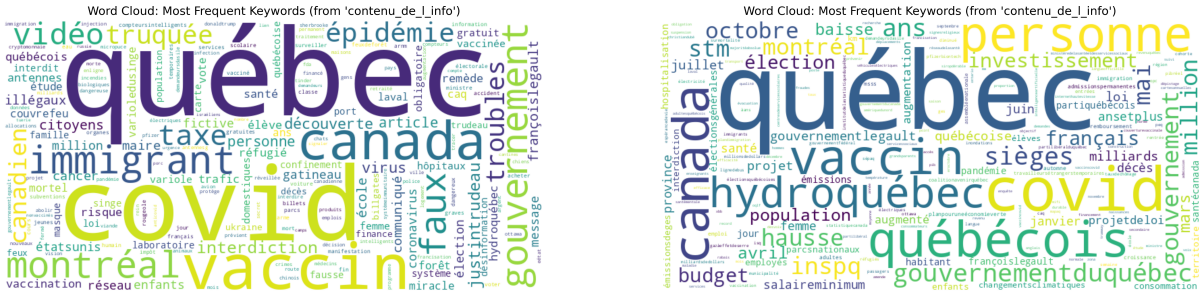
Infox-QC spans a wide range of domains, including health, immigration, politics, economy, climate, and technology, offering both thematic breadth and balance across veracity classes. The details of the thematic categories presented in the dataset, as well as their distribution, are presented in Table 3 and Figure 2, respectively. Overall, the dataset shows a clear predominance of health-related content, accounting for roughly one third of all annotations. Politics, immigration, and economic issues form additional clusters of importance, while other domains appear more marginal. This thematic diversity highlights the dataset's capacity to capture not only pandemic-related misinformation but also the broader socio-political and cultural dynamics of Québec's information ecosystem.

Figure 3 provides a word cloud based on the most frequent keywords in the collected data for both true and fake news. The true news visualization conveys official, government-related terms, such as hydroquebec, budget, vaccine, and COVID. The thematic focus is strongly related to public administration. Conversely, the Fake News visualization is characterized by more sensational and negative terms. While keywords like COVID and vaccine are also prominent, they appear alongside words that imply deceit and controversy, such as fake video, taxed, trouble, Canada, immigration, and false. These visual representations show that true news centers on institutional topics and official data, while fake news leans toward political controversy, sensationalism, and conspiracy themes.

3.4. Human–LLM Thematic Annotation Consistency

To provide an additional validation signal for the thematic annotation of main topic groups, we conducted a preliminary agreement analysis between the human-assigned theme groups and independent LLM annotators. This experiment is not intended to replace traditional inter-annotator agreement protocols with multiple human annotators and adjudication, but rather to assess whether the proposed main-group taxonomy is sufficiently well-defined to be reproducible by an automated annotator.

Concretely, each LLM was prompted to assign exactly one theme group using only the textual claim field. The models did not have access to the human theme labels. Agreement with the human labels was measured using Accuracy, Macro-F1, and Cohen's κ, a chance-corrected agreement statis-



(a) Misinformation (False Claims)

(b) Verified Information (True Claims)

Figure 3: Most frequent keywords found in collected data present in Infox-QC.

Category	Description
Health	Focuses on public health, vaccination, COVID-19, medicine, and prevention, reflecting its prominence in misinformation during the pandemic.
Others	Aggregates diverse topics like satire and celebrity claims, with potential for future refinement.
Immigration & Society	Covers asylum, migration, and social integration, often linked to political debates.
Politics & Governance	Includes elections, governance, and language policy, underscoring politicized narratives.
Economy & Work	Addresses taxation, employment, and digital economy, highlighting socioeconomic concerns.
Environment & Climate	Encompasses climate change, biodiversity, and urban planning, reflecting ecological themes.
Education & Technology	Combines education, science, and digital innovation, showing thematic overlap.
Security & Crime	Includes law enforcement, terrorism, and digital fraud, indicating security-related misinformation.
Culture & Media	Covers cultural production and media practices.
Transport & Infrastructure	Focuses on transportation and urban development.

Table 3: Main Thematic Category Distribution

tic commonly used for inter-rater reliability (Cohen, 1960; Landis and Koch, 1977). As shown in Table 2, both LLM annotators exhibit substantial agreement with the human annotations, supporting the internal consistency and clarity of the theme-group scheme. Performance is lowest for the *Others* group, which is expected given its heterogeneous nature and low support.

4. Experimental Setup

In this section, we present the evaluation setup and metrics used to assess several fake news detection models on the Infox-QC dataset. The objective is two-fold. First, we provide baseline results for Infox-QC, with a comparative analysis of the performance of traditional machine learning classifiers, BERT-based neural networks, and prompt-based large language models (LLMs), including RAG-based approaches. Additionally, we evaluate these models' vulnerability to AI-generated fake news.

4.1. Benchmark Models

We considered three classes of benchmark classifiers. The Machine learning classifiers included SVM, Logistic Regression, Decision Tree, Gradient Boosting, Neural Network, and SGDClassifier. All of these methods were trained using TF-IDF features. For the Bert-base classifier, we relied on CamemBERT-base, which is a French version of the original BERT. The prompt-Based were based on four different LLMs, that is, GPT-3.5-turbo, , GPT-4o-mini, Claude-3-Haiku, and Gemini-1.5-flash. Finally, a variation of the previous approach, namely Retrieval-augmented generation, where an external knowledge containing the training data was used, was also considered.

4.2. Ai-generated Fake News

To evaluate how well fake news detection models handle AI-generated misinformation, we created synthetic fake news samples using the GPT-4o-mini model. We chose this model because of its strong multilingual capabilities, high fluency in French, and efficiency in producing coherent and contextually consistent text, which makes it good for generating realistic disinformation. We used two different prompting strategies to produce synthetic fake news, generating 500 samples for each method to match the number of original fake news articles in the dataset. In the first method (namely, *Fake 1*), we asked the model to turn genuine news

Feature	True news	Fake news	*Fake 1*	*Fake 2*
Number of tokens	30.0060	18.6420	240.1200	463.9840
Number of types	19.6420	14.1840	128.3320	215.1120
Size of words (characters)	5.1230	5.2456	5.5587	5.4421
Type-token ratio	0.6677	0.8102	0.5495	0.4649
Number of sentences	1.0940	1.1380	7.8560	15.8700
Size of sentences (words)	20.3601	13.8682	25.7984	24.0780
Verb-to-token ratio	0.0731	0.1085	0.1107	0.1038
Noun-to-token ratio	0.2429	0.2544	0.2316	0.2229
Adjective-to-token ratio	0.0676	0.0728	0.0816	0.0743
Adverb-to-token ratio	0.0226	0.0158	0.0333	0.0388
Pronoun-to-token ratio	0.0096	0.0114	0.0245	0.0372
Stopword-to-token ratio	0.3789	0.3976	0.4418	0.4466
Misspelling-to-token ratio	0.0000	0.0000	0.0000	0.0000
Pausality	3.1740	1.2282	2.9099	3.2163
Emotiveness	0.3101	0.2680	0.3390	0.3478
Uncertainty	0.0400	0.0680	1.7600	4.3440
Non-immediacy (individual references)	0.0000	0.0005	0.0001	0.0007
Non-immediacy (group references)	0.0000	0.0006	0.0008	0.0049

Table 4: Linguistic analysis of our dataset containing average values of linguistic features for original true and fake news, and AI-generated fake news from true (*Fake 1*) and fake news (*Fake 2*).

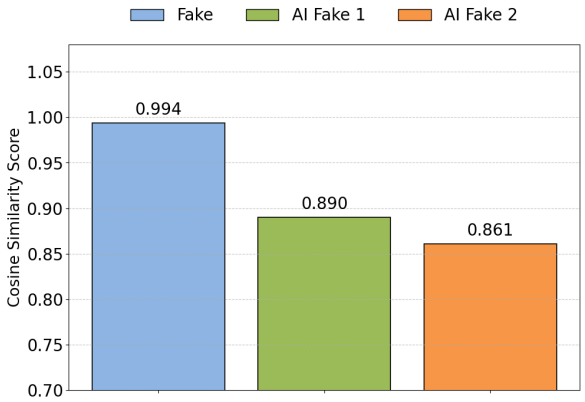


Figure 4: Cosine similarity between linguistic features extracted from true and fake news.

articles into realistic and engaging fake news stories by subtly changing factual details and using common misinformation techniques like exaggeration, emotional framing, and false attribution. In the second method, referred to as *Fake 2*, we instructed the model to improve existing fake news articles, making them more realistic and convincing by enhancing linguistic coherence and narrative plausibility while keeping the overall theme of the original fake content.

4.3. Experimental Settings

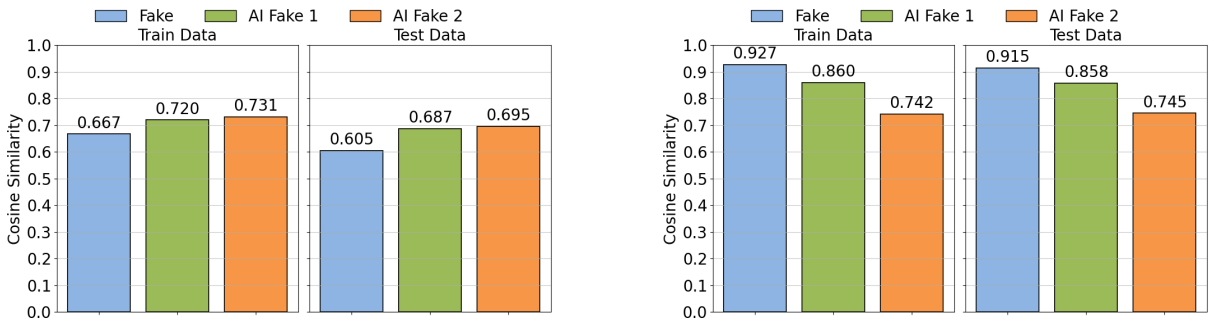
A standardized preprocessing pipeline, which included text normalization, stopword removal, lemmatization, and label harmonization, was applied to ensure clean and consistent text data. The dataset was subsequently partitioned through random sampling into training and test subsets, specifically with 70% of the data allocated for training and 30% reserved for testing. This split ensured balanced class representation and non-overlapping entries. Performance was assessed using accuracy, precision, recall, and F1-score.

5. Experimental Results

5.1. Linguistic Analysis

Following the work presented in (Silva et al., 2025), we computed NLP features based on linguistic elements, such as number of tokens, types, sentences, and verbs, among others. We also considered linguistic features proposed in (Zhou et al., 2004), including pausality, which refers to the number of pauses in text (e.g., punctuation); emotiveness, which is the expressiveness of the language based on the number of adjectives and adverbs divided by the number of nouns and verbs; uncertainty, which relates to the number of modal verbs and passive voice; and individual and group non-immediacy, which is the occurrence of first- and second-person pronouns (Silva et al., 2025). These features were computed from original true and fake news, AI-generated fake news from true news (i.e., *Fake 1*), and AI-generated fake news from fake news (i.e., *Fake 2*).

Table 4 reveals significant differences in style and structure between human-written and AI-generated text. Human-written content is relatively concise, with true news averaging around 30 tokens while fake news just 19 tokens. Note that the content generated by the LLMs is longer, reaching over 460 tokens for the samples generated from human-written fake news. This increased length in AI-



(a) TF-IDF

(b) BERT

Figure 5: Cosine similarity between true news and fake news based on embeddings representing the whole dataset.

generated text correlates with a lower Type-token ratio, suggesting more lexical repetition, and a higher stopword-to-token ratio. The most significant divergence lies in the psycholinguistic features, where both LLM categories exhibit drastically higher uncertainty scores compared to human-written articles, with *Fake 1* scoring 1.7600 and *Fake 2* scoring 4.3440. This suggests the AI models frequently use cautious language, such as modal verbs and passive voice, to express claims. Additionally, *Fake 2* shows the highest Pausality and Emotiveness scores, indicating that the AI outputs are not only much longer but also structurally and stylistically more complex and cautious than either true or human-written fake news.

AI-generated content is not more similar to true news than original fake news. It fails to replicate the key characteristics of both human categories. While the LLMs match true news in terms of sentence length and emotiveness, their outputs are vastly longer (i.e., more tokens) and contain dramatically higher uncertainty scores (e.g., 4.3440 vs. 0.0400). This dissimilarity is quantitatively confirmed by the cosine similarity provide in Figure 4, where the vector of true news features show an almost perfect match with original fake news, 0.994, but a weaker match with fake news based on true articles (i.e., *Fake 1*), 0.890, and fake news based on false articles (i.e., Fake 2), 0.861.

Figure 5 presents the cosine similarities between true news and three distinct categories of misinformation: human-written fake news, AI-generated fake news from true articles (*Fake 1*), and AI-generated fake news derived from false articles (*Fake 2*). These results are evaluated using both lexical and semantic representations, respectively represented by TF-IDF and BERT embeddings. The TF-IDF model, shown in Figure 5-a, was trained using the training set, with the test set similarities calculated by adjusting the model using only original news from the training phase. In contrast, Figure 5-b utilizes BERT embeddings to capture semantic relationships that TF-IDF may overlook. According to the TF-IDF results, AI-generated fake news (Fake 2) shows the highest lexical similarity to true news (0.731 on Train), suggesting that AI-generated content more closely resembles the word choice of true news than human-written fake news does (0.667). This creates a challenging scenario for classifiers trained solely on human-written text. However, the BERT semantic embeddings reveal a different trend; here, human-written fake news achieves the highest similarity to true news (0.927 on Train), while AI-generated Fake 2 shows the lowest semantic correlation (0.742). These findings indicate that while AI may lexically mimic true news effectively, it remains semantically distinct compared to human-authored misinformation.

5.2. Evaluation on Human-written text

Table 5 summarizes model performance on human-written fake news. The results show the RAG-based and LLM-based architectures outperforming Bert-based and traditional baselines, with the RAG system based on Gemini achieving the best overall F1-score, 0.95, followed by RAG based on GPT-4o-mini, which achieved 0.92. These models excel at identifying nuanced linguistic cues and maintaining factual consistency, benefiting from retrieval-augmented reasoning and broader contextual grounding. Within the LLM group, Claude achieved the best F1-score, 0.93, outperforming both GPT-4o-mini, 0.90, and Gemini, 0.89. Performance among the traditional ML classifiers was generally strong but below the best LLMs. CamemBERT, for instance, provides the best performance, i.e., F1-score of 0.91, among the ML models. The lowest performance was recorded by Gpt2, 0.66, and Gradient Boosting, 0.80.

	Model	Precision	Recall	F1-score
RAG	Gpt3.5	0.90	0.90	0.90
	Gpt4	0.92	0.92	0.92
	Gemini	**0.95**	**0.95**	**0.95**
	Claude	0.88	0.86	0.86
LLM	Gpt2	0.77	0.69	0.66
	Gpt-3.5	0.89	0.88	0.88
	Gpt-4o-mini	0.90	0.90	0.90
	Gemini	0.89	0.89	0.89
	Claude	0.93	0.93	0.93
ML	CamemBERT	0.91	0.91	0.91
	SVM	0.89	0.89	0.89
	Neural Network	0.88	0.88	0.88
	Decision Tree	0.82	0.82	0.82
	Logistic Regression	0.88	0.88	0.88
	Gradient Boosting	0.81	0.80	0.80
	SGDClassifier	0.88	0.88	0.88

Table 5: Performance of models trained and tested with human-written fake news.

	Model	Precision	Recall	F1-Score
RAG	RAG-Gpt3.5	0.80	0.79	0.79
	RAG-Gpt4	0.85	0.84	0.84
	RAG-Gemini	0.95	0.95	0.95
	RAG-Claude	0.93	0.93	0.93
LLM	Gpt2	0.24	0.45	0.31
	Gpt-3.5	0.81	0.76	0.77
	Gpt-4o-mini	0.90	0.89	0.89
	Gemini	0.93	0.93	0.93
	Claude	**0.96**	**0.95**	**0.95**
ML	CamemBERT	0.93	0.93	0.93
	SVM	0.53	0.51	0.43
	Neural Network	0.52	0.51	0.46
	Decision Tree	0.47	0.48	0.44
	Logistic Regression	0.66	0.60	0.56
	Gradient Boosting	0.65	0.65	0.65
	SGDClassifier	0.53	0.52	0.45

Table 6: Performance of models trained with human-written and tested with AI-generated fake news from true samples.

5.3. Evaluation on AI-Generated fake news from true news

In this experiment, we trained the models on human-written and tested on AI-generated misinformation. Note that the synthetic fake news were generated based no the first method (i.e., Fake 1), which relies on altering true news articles into a fake ones. Results in Table 6 show that most models exhibited performance decay, revealing sensitivity to domain shift. The RAG system based on Gemini and Claude maintained robust performance, F1-score of 0.95, while traditional ML models, such as SVM and Decision Tree, struggled significantly. This indicates that LLMs and hybrid models generalize better to synthetic linguistic patterns, leveraging context reasoning rather than surface lexical cues. Specifically, the LLM group demonstrated exceptional resilience, with Claude achieving the highest overall F1-score (0.95), closely followed by Gemini (0.93) and GPT-4o-mini (0.89). The performance of the RAG models was also very strong, with RAG-Gemini leading its group at 0.95 F1. The sharp performance decline was concentrated among the ML classifiers: SVM dropped to an F1-score of 0.43, and Decision Tree to 0.44, highlighting their reliance on features easily manipulated or altered by AI generation. CamemBERT, however, showed resilience for a non-LLM model, achieving an F1-score of 0.93. Gpt2 showed the worst performance, with a severe drop to 0.31 F1, suggesting a lack of robustness when faced with out-of-distribution AI text.

5.4. Evaluation on AI-Generated fake news from false news

Table 7 presents results where models trained on human-written fake news are tested on AI-generated, i.e., more realistic fake news. Results show a widespread and severe drop in performance across almost all categories, indicating a significant challenge in detecting this particularly sophisticated form of synthetic misinformation. The overall top performance comes from the LLM category, specifically Claude, which achieved the highest F1-score (0.87), followed by Gemini (0.79). This indicates that modern, advanced LLMs are the most robust against this novel domain shift. The performance of the traditional ML classifiers was poor, with most scores dropping below 0.50, highlighting their complete failure to generalize from human-written features to the more refined, realistic style of the AI-generated fake news (Neural Network at 0.35 F1, SVM at 0.36 F1). Notably, the RAG models did not perform as well as the top direct LLMs, with RAG-Gemini peaking at 0.56 F1, suggesting that retrieval augmentation might have introduced conflicting or irrelevant information. CamemBERT (0.64 F1) was the best-performing non-LLM model, still trailing the top LLMs by a significant margin. Gpt2 remained the weakest model overall, with an F1-score of 0.31.

6. Ethical Considerations

All data used in this research were collected from publicly available and verified fact-checking sources in accordance with ethical research standards. Sensitive or personally identifiable information was excluded during preprocessing to preserve privacy. Because misinformation can involve politically or socially charged topics, care was taken to avoid amplifying false narratives or creating datasets that could be misused. The dataset is released strictly for academic and educational

	Model	Precision	Recall	F1-Score
RAG	RAG-Gpt3.5	0.55	0.55	0.53
	RAG-Gpt4	0.50	0.50	0.46
	RAG-Gemini	0.73	0.61	0.56
	RAG-Claude	0.69	0.62	0.58
LLM	Gpt2	0.24	0.45	0.31
	Gpt-3.5	0.58	0.57	0.57
	Gpt-4o-mini	0.58	0.55	0.52
	Gemini	0.80	0.79	0.79
	Claude	**0.88**	**0.87**	**0.87**
ML	CamemBERT	0.72	0.67	0.64
	SVM	0.42	0.47	0.36
	Neural Network	0.37	0.45	0.35
	Decision Tree	0.46	0.47	0.42
	Logistic Regression	0.30	0.46	0.33
	Gradient Boosting	0.49	0.49	0.46
	SGDClassifier	0.40	0.47	0.36

Table 7: Performance of models trained with human-written and tested with AI-generated fake news from fake samples.

purposes. Finally, since LLMs may reflect biases in their training data, our evaluation emphasizes transparency, reproducibility, and fairness. We encourage future studies to perform systematic bias and fairness assessments before any practical deployment.

7. Future Work

Future research will extend this foundation in scale, scope, and interdisciplinarity. Expanding temporal and topical coverage will enable stronger statistical analyses and domain adaptation. A primary direction is to expand beyond textual claims and incorporate multimodal misinformation, including manipulated images, memes, and short-form videos. Contemporary misinformation increasingly operates through hybrid formats where text, visuals, and symbolic cues interact; integrating these modalities will allow Infox-QC to better reflect real-world dissemination patterns and support the development of multimodal detection systems. We also aim to progressively increase the dataset size in future releases while preserving manual verification standards and thematic consistency. Expanding the corpus will improve statistical robustness, enable more fine-grained analyses, and support stronger generalization across model families. Finally, we intend to move beyond traditional binary classification toward uncertainty quantification in misinformation detection. Rather than producing rigid true/false outputs, future models will estimate degrees of epistemic uncertainty, enabling more nuanced predictions and supporting transparent, human-in-the-loop fact-checking workflows. Emphasis will also shift toward transparency, with retrieval-augmented systems offering interpretable evidence and uncertainty-aware outputs.

8. Conclusion

We introduce a balanced, Québec-focused misinformation dataset and a comprehensive benchmarking suite spanning traditional ML, transformer fine-tuning, prompt-based LLMs, and retrieval-augmented generation. The Québec focus constitutes the dataset's principal advantage: it enables culturally and linguistically grounded analysis, yields more context-sensitive modeling, and exposes patterns that broader global corpora may obscure. Empirically, retrieval-augmented LLMs produce the best veracity performance, while classical methods remain valuable for lightweight and interpretable deployments. In addition, our exploratory human–LLM thematic agreement analysis supports the internal consistency of the proposed thematic grouping. By releasing the resource and evaluation protocols, we aim to catalyze regional research, and contribute a diagnostic benchmark for French-language misinformation studies with particular relevance to the Québec region. Ultimately, this work demonstrates that targeted, regionally informed datasets are essential building blocks for trustworthy, context-aware misinformation detection.

9. Ethical Statement and Limitations

The Québec-centered focus of this study is both a methodological strength and a defining boundary. Concentrating on one linguistic and cultural environment enables precise, context-aware modeling often lost in broader multilingual datasets, while naturally limiting direct generalization to other Francophone regions. However, several limitations shape interpretation. The current corpus of 1,000 verified claims, while balanced and representative, remains modest for large-scale pretraining or deep generalization and is therefore best suited as a diagnostic benchmark. Its exclusive focus on text currently omits multimodal misinformation—memes, manipulated visuals, or videos—that drive much online diffusion. Although factual labels are verified, thematic annotation involves subjective judgment; we provide a preliminary human–LLM agreement analysis, but future versions should include multi-annotator validation and agreement metrics. Finally, reliance on public fact-checking archives may underrepresent fleeting or under-documented content. These constraints clarify rather than diminish the dataset's value: it prioritizes cultural depth, linguistic authenticity, and interpretability over scale, positioning Infox-QC as a culturally grounded benchmark for Francophone misinformation research.

10. Bibliographical References

Mohammed Abbas Yousef, Abeer ElKorany, and Hanaa Bayomi. 2024. Fake-news detection: a survey of evaluation arabic datasets. *Social Network Analysis and Mining*, 14(1):225.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Daniel Allington, Bobby Duffy, Simon Wessely, Nayana Dhavan, and James Rubin. 2021. Health-protective behaviour, social media usage and conspiracy belief during the covid-19 public health emergency. *Psychological medicine*, 51(10):1763–1769.

Joanna M Burkhardt. 2017. *Combating fake news in the digital age*, volume 53. American Library Association Chicago, IL, USA.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Botambu Collins, Dinh Tuyen Hoang, Ngoc Thanh Nguyen, and Dosam Hwang. 2021. Trends in combating fake news on social media–a survey. *Journal of Information and Telecommunication*, 5(2):247–266.

DeepSeek-AI. 2025. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *arXiv preprint arXiv:2401.08281*.

Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. 2020. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nature human behaviour*, 4(12):1285–1293.

Benjamin Icard, François Maine, Morgane Casanova, Géraud Faye, Julien Chanson, Guillaume Gadek, Ghislain Atemezing, François Bancilhon, and Paul Égré. 2024. A multi-label dataset of french fake news: Human and machine insights. *arXiv preprint arXiv:2403.16099*.

Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. 2021. Fakebert: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Cailin O'Connor and James Owen Weatherall. 2021. Modeling how false beliefs spread. *The Routledge Handbook of Political Epistemology*, pages 203–213.

Darius Plikynas, Ieva Rizgelienė, and Gražina Korvel. 2025. Systematic review of fake news, propaganda, and disinformation: Examining authors, content, and social impact through machine learning. *IEEE Access*, 13:17583–17629.

Julie Posetti and Alice Matthews. 2018. A short guide to the history of 'fake news' and disinformation. *International Center for Journalists*, 7(2018):2018–07.

Farooq Azam Rathore and Fareeha Farooq. 2020. Information overload and infodemic in the covid-19 pandemic. *J Pak Med Assoc*, 70(5):S162–S165.

Tony Berber Sardinha. 2024. AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1):100083.

Anita Saroj and Sukomal Pal. 2020. Use of social media in crisis management: A survey. *International Journal of Disaster Risk Reduction*, 48:101584.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Renato Moraes Silva, Hazem Amamou, Lucca Baptista Silva Ferraz, Fabio Kauê Araujo da Silva, and Anderson Raymundo Avila. 2025. Fake news detection in portuguese under large language model-generated content. *Journal of the Brazilian Computer Society*, 31(1):1150–1167.

Renato Moraes Silva, Pedro Reis Pires, and Tiago A. Almeida. 2023. Incremental learning for fake news detection. *Journal of Information and Data Management*, 13(6).

Soubraylu Sivakumar, Lakshmi Sarvani Videla, T Rajesh Kumar, J Nagaraj, Shilpa Itnal, and D Haritha. 2020. Review on word2vec word embedding neural net. In *2020 international conference on smart electronics and communication (ICOSEC)*, pages 282–290. IEEE.

Steven T Smith, Edward K Kao, Erika D Mackin, Danelle C Shah, Olga Simek, and Donald B Rubin. 2021. Automatic detection of influential actors in disinformation networks. *Proceedings of the National Academy of Sciences*, 118(4):e2011216118.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

James Owen Weatherall and Cailin O'Connor. 2024. Fake news! *Philosophy Compass*, 19(6):e13005.

Haoning Xue, Xuanjun Gong, and Hannah Stevens. 2022. Covid-19 vaccine fact-checking posts on facebook: Observational study. *Journal of medical Internet research*, 24(6):e38423.

L. Zhou, J.K. Burgoon, D.P. Twitchell, T. Qin, and J.F. Nunamaker Jr. 2004. A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20(4):139–165.