
Getryt: AI misinformation detection and verification system for digital safety ¹

Olive Mbuaya Dibula 1
Birgham Young
University, United
States

Kabo Molefe 2
University of
Witswatersrand, South
Africa

Azania Makina 3
University of Cape
Town, South Africa

With
Apart Research

Abstract

The growing capabilities of artificial intelligence has enabled the large-scale generation of highly convincing misinformation, scams, impersonation attempts and synthetic media across digital platforms. In South Africa, where millions of people rely on online communication channels to access and share information, these developments have increased users' vulnerability to fraudulent content, deceptive media and AI-enabled manipulation. These risks are particularly significant among individuals with limited digital and AI literacy who may struggle to distinguish authentic content from AI-generated or manipulated information.

To address this challenge we developed Getryt, an AI-powered detection and verification tool that is designed to help users evaluate suspicious content from a variety of online sources. The system enables users to submit potentially misleading messages for analysis, where content is evaluated for indicators of misinformation, fraud and other kinds of AI-generated deception. By providing concise explanations and evidence-based assessment. In addition to flagging potentially harmful content, Getryt provides users with concise assessments, such as whether content is likely AI-generated or potentially false due to a lack of supporting evidence from official sources and directs them to trusted channels for further verification.

Evaluation across representative real-world scenarios demonstrated the prototype's effectiveness in identifying common scam patterns and misleading content while delivering rapid, actionable feedback. Our findings suggest that accessible AI-enabled deception. The project demonstrates the potential of user-centered detection systems to support safer and more trustworthy digital information environments.

¹ Research conducted at the [The Global South AI Safety Hackathon](#), June 2026

1. Introduction

The rapid advancement of artificial intelligence has transformed how information is created, distributed and consumed across digital platforms. While these technologies offer a lot of benefits, they have also made it increasingly easy to generate and spread highly convincing misinformation, impersonation schemes and fraudulent content at scale. AI-generated messages, images and text are now commonly used to create fake job advertisements, misleading health information, election-related misinformation and various financial scams. These challenges are global in nature, but they are especially critical in the African context particularly in South Africa, where digital communication platforms play a central role in everyday life.

A key risk lies in the growing sophistication of AI-generated content which is often indistinguishable from legitimate information. Malicious actors exploit this by targeting users on widely used messaging platforms, impersonating trusted institutions such as government agencies or financial aid programs (e.g NSFAS) and spreading misleading or fraudulent content directly within private conversations. This threat is amplified by limited digital and AI literacy, which makes it difficult for many individuals to critically evaluate the authenticity of the information they receive. As a result, users are increasingly vulnerable to scams, financial loss and harmful decision-making based on false information. While existing approaches to misinformation detection exist, they require users to leave their communication

To address this challenge, we developed Getryt, an AI-powered online detection and verification tool that enables users to verify information by inputting content for analysis. The system analyzes user-submitted or forwarded messages, identifies potentially AI-generated or misleading content, and flags risks such as fraudulent job advertisements, scam messages and misinformation campaigns. Importantly, it goes beyond detection by redirecting users to verified and trusted sources, such as official government or institutional websites, enabling immediate validation and informed decision-making. This approach transforms misinformation detection into a practical user-centered tool for digital safety.

Key contributions:

Our work advances the current state of misinformation detection through the following key contributions:

- **Accessible AI detection systems** - We introduce a user-friendly online system that enables real-time verification of suspicious content without requiring technical expertise, making misinformation detection more accessible to everyday users

- **Localized and context-aware risk detection** - Our approach is specifically tailored to the South African context, allowing for the identification of region-specific scams such as fake government job postings, NSFAS fraud and locally relevant misinformation
- **Actionable Verification through trusted redirection** - Unlike traditional tools that only flag suspicious content, our system provides actionable outcomes by directing users to verified official sources, bridging the gap between detection and informed decision-making

2. Related Work

The detection of misinformation and AI-generated content has been widely explored across natural language processing (NLP), machine learning and digital media forensics. A substantial body of research focuses on automated fake news detection using machine learning and deep learning techniques. Recent studies show that transformer-based models, such as BERT, have significantly improved classification accuracy by capturing contextual relationships in text, however these approaches are typically evaluated on curated datasets and may not generalize effectively to real-world environments where misinformation evolves rapidly (Alshuwaier & Alsulaiman, 2025). From an NLP perspective, misinformation detection remains an open and complex challenge due to the dynamic nature of language, cultural variation and domain specific usage. Prior work emphasizes that detection systems often struggle to adapt across different regions, particularly those with underrepresented data, such as parts of Africa. This limitation reduces the effectiveness of global models when applied to localized misinformation patterns, including region-specific scams and impersonation tactics (Su et al., 2020). Research on AI-generated text detection has further explored the challenge of distinguishing human-written content from machine generated text. Studies such as ‘On the possibilities of AI-generated text detection’ demonstrate that while detection is achievable, it becomes increasingly difficult as generated text becomes more human-like (Chakraborty et al., 2023). Additionally empirical evaluations of existing detecting tools reveal that many systems suffer from limited reliability and can be easily bypassed through paraphrasing or minor modifications. These findings highlight the challenges of deploying robust detection in real-world scenarios (Weber-Wulff et al., 2023)

Beyond text-based misinformation, the broader field of synthetic media detection, including deepfakes has also been extensively studied. Studies identify key issues such as model robustness, interpretability and reliability in practical applications indicating that many current systems are not yet fully suited for everyday use (Wang et al., 2022)

In parallel, research on misinformation spread in the Global South highlights the importance of messaging platforms such as Whatsapp. Studies conducted in South Africa demonstrate that misinformation is frequently disseminated through these platforms and that interventions like fact-checking delivered via Whatsapp can improve users’ ability to identify false information, Although such approaches typically rely on periodic information delivery rather than real time, user-driven verification (Bowles et al., 2020).

Despite these advances, existing approaches face several critical limitations:

- Lack of platform integration - most tools are standalone systems that are not integrated with everyday communication platforms
- Limited localization- many models fail to capture region-specific misinformation and fraud patterns
- Limited actionability- current tools typically classify or flag content without guiding users toward verified information

Our work addresses these gaps by introducing an AI-powered detection tool designed for real-time user driven analysis of suspicious content. Unlike prior approaches, our system is context-aware and localized, focusing on misinformation relevant to South Africa. Furthermore, it goes beyond detection by providing actionable verification through trusted redirection, transforming passive analysis into active user protection.

Our method is particularly advantageous when:

- (a) users encounter suspicious content in digital communication environments
- (b) Users have limited digital literacy or AI literacy and require a simple interface
- (c) Misinformation is context-specific and not well represented by global datasets

Overall, this work demonstrates that:

- AI detection systems improve accessibility when designed for real-world use
- Localized, context-aware detection improves effectiveness and efficiency
- Combining detection with trusted redirection provides a more complete and practical solution to misinformation

3. Methods

System overview: We developed an AI detection system that enables users to verify suspicious messages in real time. The system follows a retrieve-and-match approach where user inputs are mapped to a curated database of verified claims rather than generating new information. This design reduces hallucination risks and ensures outputs remain grounded in trusted sources.

Although the full system is designed for integration into messaging environments, the current prototype is implemented as a web-based demonstration tool due to resource and API constraints

Architecture and implementation: The system is built using [Node.js](#) backend responsible for API integration message routing and verification logic. A lightweight frontend (HTML,CSS and JavaScript) support a web-based demo interface for user interaction and demonstration purposes. The backend manages user input, queries the verification pipeline and returns structured responses based on matched or retrieved information

Tools and Technologies: Backend - [Node.js](#) ; Frontend - HTML,CSS, JavaScript ; Development tools - GitHub, Powershell ; AI Support - Google Gemini (Prompting and prototyping)

Design Choices: The system is inspired by messaging-based interaction models namely; Meta AI and Twilio but differs by prioritizing verification over generation. This ensures higher trustworthiness, especially when it comes to issues such as scams and misinformation

Challenges and Iterations: Multiple LLM APIs eg. OpenRouter, introduced inconsistency. Fully generative outputs reduced the reliability due to hallucination risk and resource heavy setups were limited by API credits and tooling constraints. As a result the system was refined toward a lightweight, retrieval-based design that prioritizes accuracy, stability and feasibility within prototype constraints.

4. Results

1. The system was evaluated using a small set of representative real-world scenarios, including fraudulent job advertisements, financial scams, impersonation messages, and general non-suspicious text. The aim was to assess whether Getryt could correctly identify suspicious content and provide reliable verification feedback.

Key Findings:

- The system accurately identified most structured scam messages (e.g. fake job adverts and financial scams). This shows that the retrieval-based approach is effective for detecting repeated and pattern-based misinformation.
 - Inputs not present in the curated knowledge base were returned as “Not yet verified.” This improves safety by avoiding false claims but limits coverage for new or evolving scams.
 - The system produced fast responses across all test cases. The lightweight architecture supports real-time use.
-
2. Performance was consistent across structured scam patterns and did not change significantly with minor wording variations. However, performance decreased for unseen or novel inputs due to the limited size of the curated knowledge base.

3. PROTOTYPE PERFORMANCE OVERVIEW:

Category	Accuracy (qualitative)	Outcome
Job scam messages	High	Mostly correctly flagged
Financial scams	High	Correctly identified

Impersonation texts	Moderate–High	Mostly detected
Normal messages	High	Correctly not flagged

5. Discussion and Limitations

The results of this work highlight the importance of delivering AI safety solutions in ways that are accessible, practical and integrated with everyday user behaviour. Our approach demonstrates that integrating misinformation and verification tools within existing communication channels can improve usability and increase the likelihood of real-world adoption. Rather than requiring users to actively seek out fact-checking services, the system enables real-time, in-context decision making within the environments where misinformation is most commonly encountered and shared.

Our results suggest that structured and context-specific misinformation, such as fraudulent job advertisements, financial scams, and impersonation attempts can be identified more reliably due to recurring linguistic, stylistic and formatting patterns. This finding highlights the value of incorporating localized and domain-specific knowledge into AI safety systems, particularly in regions where misinformation often takes forms that are underrepresented in global datasets and detection frameworks.

It is also important to consider the role of actionable feedback in strengthening digital resilience. By not only flagging suspicious content but also directing users to trusted and authoritative sources, the system moves beyond passive detection and supports informed decision-making. This approach helps bridge the gap between identifying potentially misleading information and enabling users to independently verify its accuracy. From an AI safety perspective, this project demonstrates that mitigating AI-enabled deception requires more than improvements in model performance alone. Effective interventions must also be human-centered, context-aware and accessible to diverse user populations. As generative AI continues to increase the volume and sophistication of misleading content, solutions that combine detection, education and verification within widely adopted platforms may play an important role in reducing misinformation harms and strengthening the resilience of digital communities.

Limitations

The system relies on a limited curated dataset of verified claims, which restricts its ability to detect novel or previously unseen misinformation. Messages that do not match known patterns

are returned as “Not yet verified”. Additionally, the system is vulnerable to evolving adversarial behaviour, as malicious actors may adapt their messaging strategies through paraphrasing or combining truthful and misleading information to bypass detection. Finally, due to hackathon time constraints the system has not been evaluated on large-scale datasets or across diverse real-world conditions and currently lacks features such as multilingual support and extensive performance benchmarking. As a result, the current implementation should be understood as a prototype demonstrating feasibility rather than a fully ready solution.

Future Work

A key step is the development of a larger, continuously updated knowledge base that incorporates verified information from trusted sources such as government agencies and educational institutions. Expanding the evidence base would improve coverage and enable the system to respond more effectively to emerging misinformation and fraud patterns. Future development should also focus on multimodal detection capabilities, allowing the system to analyze images, voicenotes, videos and other forms of media increasingly used in AI-enabled scams. Expanding support for South Africa’s diverse languages would improve accessibility and ensure the tool can serve a larger number of people in the country. Finally the underlying framework could be adapted for use across additional messaging and social media platforms. Future implementations could integrate with various other platforms enabling more people to access this detection tool through channels they already use. By supporting a wider range of users, platforms and regional contexts future iterations of the system could contribute to more scalable and globally relevant approaches to combating misinformation, scams and promoting safer online ecosystems.

6. Conclusion

This project presented ‘Getryt’, an AI detection and verification system designed to address the growing challenge of misinformation, scams and AI-enabled deception. The system provides users with an accessible and convenient way to evaluate suspicious information and verify claims through trusted sources. The retrieval-based approach improves reliability by grounding responses in curated evidence, reducing the risk of inaccurate or unsupported outputs.

Overall, the project demonstrates how AI safety tools can be designed for practical use by focusing on accessibility and usability. Getryt offers a practical approach to combating misinformation in contexts where digital and AI literacy vary significantly. Although still a prototype, it highlights how accessible verification tools can help users navigate an increasingly complex digital information landscape.

Code and Data

Code repository: <https://github.com/olive497/getRyt>

Data/Datasets: [N/A]

Video walk

through:<https://1drv.ms/v/c/2fb173978d3c2906/IQApnuMdkTXNTKOVN1hMxFVSAUiDpPj-XwbNO46kal7TeM?e=mS0s3e>

Hugging face space: <https://huggingface.co/spaces/KaboKableMolefe/getryt-demo>

References

Alshuwaier, F. A., & Alsulaiman, F. A. (2025). Fake news detection using machine learning and Deep Learning Algorithms: A comprehensive review and future perspectives. *Computers*, 14(9), 394.

<https://doi.org/10.3390/computers14090394>

Bowles, J., Croke, K., Larreguy, H., Liu, S., & Marshall, J. (2020–2022). Combating misinformation using fact-checking via WhatsApp in South Africa. Abdul Latif Jameel Poverty Action Lab (J-PAL).

<https://www.povertyactionlab.org/evaluation/combating-misinformation-using-fact-checking-whatsapp-south-africa>

Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., & Huang, F. (2023). On the possibilities of ai-generated text detection. *arXiv preprint arXiv:2304.04736*.

<https://doi.org/10.48550/arXiv.2304.04736>

Su, Q., Wan, M., Liu, X., & Huang, C. R. (2020). Motivations, methods and metrics of misinformation detection: an NLP perspective. *Natural Language Processing Research*, 1(1), 1-13.

<https://doi.org/10.2991/nlpr.d.200522.001>

Wang, T., Liao, X., Chow, K. P., Lin, X., & Wang, Y. (2024). Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys*, 57(3), 1-35.

<https://doi.org/10.1145/3699710>

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1), 1-39.

<https://doi.org/10.48550/arXiv.2306.15666>

LLM Usage Statement

We used large language models (including Google Gemini ,Chat GPT,Groq,Claude code and Open code) to assist with coding support, prompt design and structuring both the backend system and this report. System functionality, outputs and results were implemented, tested and verified by the team. We have also ensured that all claims and conclusions accurately reflect the behaviour of the developed system