Check for updates

METHOD ARTICLE

REVISED Detecting FIMI: A methodological framework for testing OSINT tools in TTPs detection

[version 2; peer review: 1 approved, 3 approved with reservations]

Konstantinos Margaros[1], Theodoros Paparrigopoulos[1], Georgios Giataganas[1], Orfeas Filippopoulos[1], Nikolaos Georgiou[2], Oleksandr Manzhai[3], Panagiota Benekou[1]

[1]Center for Security Studies (KEMEA), Athens, 10177, Greece
[2]Hellenic Police, Athens, Greece
[3]Kharkiv National University of Internal Affairs, Kharkiv, Ukraine

v2 **First published:** 09 May 2026, **6**:157
https://doi.org/10.12688/openreseurope.21702.1
Latest published: 08 Aug 2026, **6**:157
https://doi.org/10.12688/openreseurope.21702.2

Abstract

Background

Foreign Information Manipulation and Interference (FIMI) represents a growing hybrid threat targeting democratic processes, social cohesion, and diaspora communities. Detecting it requires identifying underlying tactics, techniques, and procedures (TTPs) described in frameworks such as DISARM, AMITT, and MITRE ATT&CK. Although many open-source intelligence (OSINT) tools exist, their evaluation remains fragmented and lacks methodological standardisation. This article presents a framework for testing OSINT tools for TTP detection, developed within the Horizon Europe project RESONANT.

Methods

The methodology integrates three components: a structured multi-source collection of OSINT tools informed by academic, operational, and grey literature; a controlled, mixed-source evaluation corpus of project-prepared examples, public reference material, and selected third-party datasets across text, multimedia, network and infrastructure analysis; and controlled testing by law enforcement partners from Greece and Ukraine. Tools were classified into eight categories and evaluated using predefined indicators of accuracy, usability, interoperability, coverage of adversarial behaviours, and sustainability.

Open Peer Review

Approval Status ? ? ✓ ?

	1	2	3	4
version 2 (revision) 08 Aug 2026		? view	✓ view	
version 1 09 May 2026	? view	✗ view	? view	? view

1. **Kokil Jaidka**, National University of Singapore, Singapore, Singapore
2. **Giovanni Spitale**, IBME, Universitat Zurich, Zurich, Switzerland
3. **Marius Dragomir**, University of Santiago de Compostela, Santiago de Compostela, Spain
4. **Veysel Bozkurt**, Istanbul University, Istanbul, Turkey

Any reports and responses or comments on the article can be found at the end of the article.

Results

Several content-verification and multimedia-forensics tools performed well on the retained material, and infrastructure-level tools gave useful indicators; these are time-bound descriptive observations, not general rankings. Exact case-level replication is limited, as the package does not preserve the full CT-A/CT-B mapping, run-level outputs, denominators, or configurations. Social media analytics and bot-detection tools were least stable under platform API restrictions. Shortcomings recurred in multilingual processing, interoperability, explainability, and integration into threat-intelligence formats.

Conclusions

The study offers a transferable methodology for evaluating OSINT tools in the context of FIMI TTP detection. Combining a mixed-source corpus, structured testing, practitioner validation, and publication of the available material, it advances methodological transparency while clarifying the limits of both automated outputs and the retained replication record. The proposed tool suite is a layered decision-support baseline, not a verdict-producing system. Future work should expand multilingual and out-of-distribution testing, operationalise bias and explainability metrics, preserve run-level records, and pursue lawful access to platform data.

Keywords
FIMI, disinformation, OSINT, TTP detection, DISARM, AMITT, threat intelligence, tool evaluation



This article is included in the Horizon Europe gateway.

Corresponding author: Konstantinos Margaros (k.margaros@kemea-research.gr)

Author roles: Margaros K: Conceptualization, Methodology, Project Administration, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Paparrigopoulos T**: Conceptualization, Formal Analysis, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Giataganas G**: Data Curation, Formal Analysis, Investigation, Validation, Writing – Original Draft Preparation; **Filippopoulos O**: Data Curation, Formal Analysis, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Georgiou N**: Investigation, Validation; **Manzhai O**: Investigation, Validation; **Benekou P**: Methodology, Project Administration, Supervision

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No [101132439] (RESONANT).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2026 Margaros K *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Margaros K, Paparrigopoulos T, Giataganas G *et al.* **Detecting FIMI: A methodological framework for testing OSINT tools in TTPs detection [version 2; peer review: 1 approved, 3 approved with reservations]** Open Research Europe 2026, **6**:157 https://doi.org/10.12688/openreseurope.21702.2

First published: 09 May 2026, **6**:157 https://doi.org/10.12688/openreseurope.21702.1

REVISED **Amendments from Version 1**

This version responds to the three peer-review reports on version 1. The revisions are substantive but do not change the study's design or core conclusions.

First, the evaluation corpus is now described accurately as a controlled, mixed-source collection — combining project-prepared examples with public and third-party reference material — rather than as fully synthetic data. Its provenance, privacy status and licensing are documented, and the complete evaluation input corpus, together with a README, has now been deposited on Zenodo; the previous deposit contained only the metrics document and the documentation template.

Second, the tool-performance figures are reframed as time-bound descriptive observations from specific tools, datasets and operational settings, not as general performance rankings or a reproducible benchmark. We state explicitly that the package does not preserve the full CT-A/CT-B mapping, run-level outputs, denominators or model configurations, and we have withdrawn the systematic-review framing.

Third, we address generative-AI tools directly: we add a caveat that high apparent accuracy on publicly documented content may reflect training-data recognition rather than independent verification, and we discuss the documented source-framing bias of large language models, qualifying the GPT-4o recommendation with operational safeguards.

Fourth, we distinguish legal-compliance ethics from epistemic ethics, introduce a decision protocol governing when verdict-producing tools may inform investigative action, and theorise the limits of content-dependent detection relative to content-agnostic approaches.

Finally, we reconcile the reported indicators with the deposited seven-metric scale and add author contributions and funding information. The results themselves are otherwise unchanged.

Any further responses from the reviewers can be found at the end of the article

1. Introduction

FIMI has emerged as a strategic hybrid threat undermining democratic resilience, electoral integrity, and societal cohesion. Recent reports by the European External Action Service[1,2] identify FIMI as a growing transnational phenomenon that exploits information environments to manipulate perceptions and disrupt decision-making processes. Similar concerns are echoed in the Canadian Special Report on FIMI, which highlights the vulnerabilities of democratic institutions and the role of hostile actors in exploiting information ecosystems.[3] Beyond the Euro-Atlantic sphere, the spread of manipulative content in fragile contexts, including Africa and the Middle East, further illustrates the global salience of these activities.

To study and counter FIMI effectively, the focus must extend beyond narratives to the TTPs used by threat actors. TTPs provide structured analytical entry points for identifying patterns of disinformation and information suppression, enabling the detection of recurrent methods across incidents. Frameworks such as the DISARM Navigator and AMITT (Adversarial Misinformation and Influence Tactics and Techniques) have been developed to classify and systematize such activities, mirroring the structured approach of MITRE ATT&CK in cybersecurity.[4] These taxonomies enable practitioners to map incidents onto standardised phases of disinformation operations, such as content creation, amplification, suppression, and mobilization, facilitating cross-case comparison and countermeasure design.

A terminological clarification is warranted at the outset. FIMI is used in this article as an umbrella term, but within the detection context two of the phenomena it subsumes carry distinct operational implications: misinformation (false content circulated without deliberate intent to deceive) and disinformation (false content deployed with strategic intent). The TTP-based approach adopted here is oriented primarily towards disinformation, since it focuses on how operations are conducted; content-level fact-checking tools, by contrast, assess whether content is false and are largely intent-agnostic. Where relevant, the analysis therefore distinguishes which layer of the tool suite addresses which type of epistemic harm, as these map differently onto attribution procedures and legal-responsibility frameworks.

Despite these advances, the detection ecosystem remains fragmented. Numerous OSINT and social media analytics tools exist, ranging from event detection systems[5] and topic discovery through social network analysis[6] to bot detection[7] and multimedia verification. Yet, these tools are rarely interoperable and often lack systematic integration with structured TTP frameworks. In addition, significant challenges persist in terms of multilingual coverage, real-time detection, and the incorporation of detection outputs into attribution processes. The problem is compounded by the ethical and legal constraints of data collection, including privacy obligations and the need for data protection impact assessments (DPIAs) in high-risk processing.[8]

Attribution further illustrates the centrality of detection capabilities. As Sabir[9] argues, credible deterrence in cyberspace hinges on reliable attribution; yet anonymity, proxies, and false flag operations frequently undermine accountability. Tsagourias and Farrell[10] distinguish between technical, political, and legal attribution, stressing that technical findings, based on indicators such as infrastructure, malware, and tradecraft, rarely translate directly into state responsibility under international law. Tools for TTP detection therefore play a pivotal but partial role: they enable technical attribution by producing observables and indicators that can feed into higher levels of attribution, but they cannot independently resolve responsibility gaps.

Against this backdrop, the Horizon Europe project RESONANT seeks to strengthen the detection and analysis of FIMI operations through an integrated methodological framework. RESONANT provides the first structured state-of-the-art review of tools for TTP detection, classifying them by function and mapping their coverage against established frameworks (DISARM, AMITT, MITRE ATT&CK). The purpose of this article is to present the results of this review in a scientific format, situating them within the broader academic and policy literature. By consolidating a fragmented tool landscape and linking it with structured taxonomies, the study contributes to building a methodological foundation for testing and validation activities in RESONANT and beyond. The work is not presented as a formal structured review in the PRISMA sense.

2. Background and conceptual framework

2.1 TTP frameworks

The concept of **TTPs** originates from cybersecurity and military doctrine, where it is used to capture the observable actions of adversaries across different stages of an operation. In the context of FIMI, TTPs provide a structured way of describing and classifying how disinformation and information suppression campaigns are conducted, from strategic planning to tactical dissemination. By focusing on the *"how"* rather than the *"what"* of disinformation, TTP frameworks enable analysts to identify recurring patterns, compare incidents, and design countermeasures.[4]

Several frameworks have been adapted to the information environment. The DISARM Navigator is one of the most widely used taxonomies for describing disinformation incidents. It structures cases across dimensions such as actor types, objectives, channels, languages, audiences, and tactics, techniques, and procedures. Its modular design supports both researchers and practitioners by enabling incident coding, cross-case comparison, and the design of countermeasures. DISARM has been adopted by the EEAS and StratCom units for structured incident reporting.[1]

Complementing this, the Adversarial Misinformation and Influence Tactics and Techniques (AMITT) framework extends the MITRE ATT&CK approach to the domain of information manipulation. It provides a comprehensive catalogue of adversarial behaviors in the information space, organised into operational phases such as planning, seeding, amplification, and suppression. Each phase is linked with specific techniques (e.g., *create fake experts, distort facts, coordinate amplification*) and mapped to potential countermeasures.[11,12] This makes AMITT particularly suitable for operationalising TTP detection, as it provides direct linkages between observed behaviors and recommended responses.

The MITRE ATT&CK framework itself, originally designed for adversary emulation in cybersecurity, remains influential in shaping the terminology and structure of TTP-based approaches. Its emphasis on standardized categorisation and machine-readable formats (e.g., STIX 2.0) has informed the development of both AMITT and DISARM. In practice, many threat intelligence platforms (e.g., OpenCTI, MISP) already integrate ATT&CK, making interoperability with AMITT and DISARM both technically feasible and desirable.[13]

Together, these frameworks provide the theoretical and methodological backbone for analysing FIMI operations. They enable a shift from narrative-centric approaches to indicator- and procedure-based detection, which is critical for technical attribution. However, as highlighted by Tsagourias and Farrell,[10] while such frameworks strengthen the technical dimension of attribution, their outputs must be complemented by political and legal processes to translate into accountability. The mapping of tools in RESONANT is therefore explicitly aligned with DISARM, AMITT and ATT&CK to ensure analytical consistency, operational relevance and future interoperability.

2.2 Social media manipulation

The diffusion of manipulative content across social media platforms is a defining feature of FIMI. Social media systems amplify adversarial tactics by lowering the barriers of entry for influence operations and enabling rapid, large-scale dissemination of content.[14] The unique affordances of social networks, interactivity, algorithmic curation, and virality, create environments where malicious actors can strategically exploit platform dynamics to distort information ecosystems.[15]

One of the most studied vectors of manipulation involves the use of automated accounts or social bots. These entities are designed to mimic human behavior, artificially inflating the visibility of certain narratives and creating false perceptions of consensus. Ferrara et al.[7] demonstrated that social bots can coordinate content amplification, hijack trending topics, and foster the illusion of grassroots mobilization, a phenomenon often described as *astroturfing*. Their findings highlight how bot-driven activity alters both the speed and scale of disinformation diffusion.

Beyond automation, manipulation also relies on the dynamics of rumor propagation. Qin et al.[16] proposed a multi-layered communication model illustrating how rumors spread and stop across social media networks. The model shows that actors strategically seed narratives within clusters of susceptible users, which can escalate to broader networks through re-sharing, while counter-rumors or debunking interventions act as dampening mechanisms. This illustrates the interplay between adversarial tactics and countermeasures in shaping the trajectory of contested information.

Another vector is the use of soft propaganda techniques. Mattingly and Yao[17] argue that persuasion in authoritarian and semi-authoritarian contexts increasingly relies on subtle framing, selective omission, and positive reinforcement rather than overt censorship. By embedding messages into entertainment formats or everyday online discourse, propagandists achieve persuasion with reduced resistance. This aligns with earlier observations that effective manipulation often combines truthful elements with misleading or distorted framing, making detection and classification more complex.

Empirical studies have further documented manipulation through case-based evidence. The *Media Manipulation Casebook* catalogues incidents where actors have exploited platform algorithms, influencer networks, and cross-platform coordination to spread disinformation. Examples range from conspiracy amplification during public health crises to adversarial exploitation of geopolitical conflicts. Such cases demonstrate that manipulation is rarely confined to a single tactic or platform; rather, it is characterized by hybridisation, opportunism, and adaptability.

Taken together, these insights underscore that social media manipulation operates through a combination of technical automation, communicative dynamics, and narrative persuasion. For detection purposes, this complexity necessitates the integration of multiple methodological approaches, including network analysis, rumor detection algorithms, bot identification, and content classification. Within RESONANT, these insights are used to inform the mapping of tools capable of addressing manipulation at different layers, technical, social, and cognitive.

2.3 OSINT evolution: Trends and security implications

OSINT has become a cornerstone of contemporary approaches to detecting FIMI. Initially conceptualised as a complementary discipline to traditional intelligence, OSINT has evolved into a primary source of indicators and observables for mapping adversarial TTPs. Its evolution reflects both technological advances and the operational needs of law enforcement, civil society, and security institutions.

Several studies highlight the trajectory of OSINT from manual collection and verification to increasingly automated, AI-enabled workflows. Dokman and Ivanjko[18] identify key challenges and trends, including data overload, source validation, and the growing reliance on social media platforms as primary arenas for intelligence collection. They emphasise that while OSINT offers opportunities for transparency and accountability, its reliability depends on methodological rigour and the ability to distinguish authentic from manipulated signals.

The role of OSINT in security contexts has expanded significantly with the proliferation of digital traceability. Hwang et al.[19] outline how the use of OSINT for cybersecurity and influence operation detection has become institutionalised in both public and private sectors. They describe OSINT as simultaneously a security enabler and a risk factor, since adversaries can also exploit open data to profile targets, craft tailored disinformation, or conduct reconnaissance for hybrid operations. This dual-use character requires refined operational standards and ethical safeguards.

At the EU level, institutional frameworks increasingly promote OSINT as a structured component of FIMI monitoring. The EEAS OSINT Guidelines for FIMI Detection[2] set methodological standards for incident documentation, indicator coding, and interoperability with frameworks such as DISARM and AMITT. These guidelines stress the importance of cross-platform monitoring, structured reporting, and alignment with legal and ethical standards, including the General Data Protection Regulation (GDPR).

The implications of these developments are twofold. First, OSINT contributes directly to technical attribution, supporting the provision of evidence such as metadata, network traces, and content analysis that can support the identification of adversarial actors and methods. Second, OSINT serves a democratic accountability function, enabling independent verification of manipulative campaigns by journalists, researchers, and civil society organisations. However, the

increasing reliance on OSINT raises critical issues around data protection, source protection, and misinformation contamination. As Demetzou[8] argues, accountability in high-risk data processing requires systematic Data Protection Impact Assessments (DPIAs) to ensure compliance and mitigate risks of overreach.

In the context of RESONANT, OSINT is recognised as a central detection layer for mapping TTPs, but its effectiveness depends on the interoperability of tools, the robustness of methodologies, and adherence to ethical standards. This dual imperative, operational efficacy and accountability, frames OSINT as both a technical and normative cornerstone of contemporary FIMI detection.

3. Methodology

The methodology applied in RESONANT was designed to provide a structured and reproducible process for testing open-source tools relevant to the detection of TTPs associated with FIMI. The process was inspired by the Waterfall model used in software engineering, ensuring a step-by-step workflow where each stage builds upon the outputs of the previous one. This approach enabled systematic evaluation while maintaining transparency and accountability. Traceability here describes the planned workflow. It does not imply that a complete run-level replication record has been preserved for every historical test.

The reference to the Waterfall model is used here only as an organising metaphor for this sequential, traceable workflow (with each stage (requirements definition, dataset preparation, execution, and analysis) producing a defined output consumed by the next) and does not imply the adoption of software-engineering process controls beyond this structuring principle.

3.1 Research requirements and tool categorisation

The first step in the RESONANT methodology was the identification of research requirements, which determined both the scope of the review and the categories of tools to be examined. Rather than approaching tool selection ad hoc, the process was anchored in established taxonomies such as DISARM, AMITT, and MITRE ATT&CK, which systematise adversarial behaviours in information operations. Aligning tool functions with these frameworks ensured that the review addressed the full spectrum of TTPs relevant to FIMI detection, from content creation and manipulation to amplification, suppression, and attribution.[18,20]

To complement this conceptual anchoring, consultations with practitioner partners, including the Hellenic Police (HP) and the Kharkiv National University of Internal Affairs/Ministry of Internal Affairs of Ukraine (KhNUIA/MoIA), were conducted. Their input helped refine the selection by highlighting which tools were most relevant to investigative workflows and law enforcement priorities. This combination of framework-driven classification and practitioner validation distinguished the methodology from purely academic reviews, embedding operational relevance into the research design.[21,22]

As a result, tools were grouped into eight functional categories: (a) fact-checking platforms, (b) website credibility assessment tools, (c) bot detection tools, (d) image and video analysis tools, (e) geo-IP verification tools, (f) phone-related tools, (g) social media analysis tools, and (h) email verification tools. Together, these categories reflect three complementary layers of detection: content-level verification (fact-checking, multimedia forensics), platform-level monitoring (social media and bot analysis), and infrastructure-level attribution (IP, phone, and email verification). This layered approach provided the foundation for structured testing and facilitated the identification of complementarities and gaps across the detection ecosystem.[23–25]

To ensure transparency and traceability, a structured selection protocol was applied. Candidate tools were identified from four source types: (i) peer-reviewed academic literature, (ii) grey literature and technical reports, (iii) the outputs and tool repositories of EU-funded projects, and (iv) practitioner recommendations from the participating law enforcement partners. RESONANT Deliverable D2.1 and the available supporting documentation record the tool inventory and access information to the extent preserved. Tools were included where they (a) were openly accessible or testable without proprietary-licensing barriers during the review period, (b) addressed at least one adversarial behaviour catalogued in DISARM, AMITT, or MITRE ATT&CK, and (c) produced outputs interpretable within an investigative workflow. Tools were excluded where they were discontinued, region-locked, inaccessible during testing, or duplicative of a better-documented equivalent. Because exact database search strings, search dates and a PRISMA-style flow record were not retained, this process is described as a structured multi-source mapping rather than a formal systematic review.

3.2 Dataset preparation and testing environment

A central methodological challenge in evaluating detection tools was the need to replicate realistic FIMI scenarios without relying on live operational data, which would have raised ethical, legal, and security concerns.

A controlled evaluation corpus was therefore assembled across text, multimedia, social-media and infrastructure layers. The retained corpus is mixed-source: it includes project-prepared illustrative records, labelled text and bot-detection material, public website and infrastructure reference data, historical email and phone-related inputs, and links to external multimedia and documented disinformation cases. Not all retained material was generated in-house or is synthetic. File-level contents, record counts, available labels, source links and limitations are documented in the README deposited with the data package.[35]

The retained corpus preserves ground truth only where it is explicitly encoded in the source material: file-level fake/real classification for the text collections, a row-level binary Bot Label for the bot-detection data, and FACT/FALSE indications for the link-based scenarios where recorded. Other files serve primarily as reference or lookup inputs and do not contain a uniform case-level outcome label. The archive does not preserve a consistent difficulty scale, case-level DISARM/AMITT annotation, complete CT-A/CT-B assignment, or formal label-adjudication record across all categories; these fields were not reconstructed retrospectively. No formal inter-rater reliability coefficient was computed. These limitations constrain the interpretation of the reported accuracy percentages. In addition, because parts of the corpus derive from or link to publicly documented material, large language models may recognise previously encountered content rather than independently verify it. The evaluation material also cannot fully reproduce the novelty and unpredictability of live operations.

The controlled corpus supported within-team comparisons because, where required by the historical test design, tools within a given category were exposed to the same selected inputs. However, the corpus contains public and third-party source material as well as project-prepared examples, and some retained files include publicly available identifiers. Ethical and data-protection safeguards therefore depend on source-specific handling, proportionality and compliance with the applicable terms of use rather than on a claim that every item is synthetic. Because complete run-level mappings are unavailable, the deposit supports inspection of the available input pool but not exact independent recomputation of every result.[18,20]

A dedicated, controlled-access environment was used by project partners to execute the historical tests and record observations about tool behaviour, usability and reliability. The workflow was designed to capture successful detections, errors and operational observations. Nevertheless, the publication package does not contain the complete historical run-level logs or output files needed to reconstruct every execution. The deposited material should therefore be understood as an available input and documentation package that improves transparency and transferability, not as a complete computational replication archive. The sequential design was inspired by waterfall-derived approaches that emphasise stepwise validation and traceability.[26–30]

3.3 Execution of testing

The third stage involved structured execution by cybercrime specialists from HP and the KhNUIA/MoIA, combining operational expertise with the controlled evaluation workflow. Scenario selection was informed by adversarial behaviours described in AMITT and DISARM.[18,20] Within each team and category, tools were applied to the same selected inputs where a direct within-team comparison was intended. Evaluators also considered whether outputs such as classifications, heatmaps and anomaly indicators were understandable and operationally useful, consistent with the principle that actionable intelligence depends on clarity as well as technical performance.[22,24]

Tool failures, access restrictions and qualitative observations informed the subsequent analysis. However, the complete historical execution trace was not retained in the deposited package. Consequently, the process is transferable at the level of its staged workflow and evaluation dimensions, while exact numerical replication of every historical run is not possible from the available archive alone.

A clarification of the testing design is warranted, since it governs how the results in Section 5 may be read. Within each cyberteam, tools belonging to a given category were applied to the same selected evaluation material where a within-team comparison was intended. However, the two cyberteams (CT-A, Hellenic Police, and CT-B, Ukrainian Police) operated on separate selections under their respective operational conditions. Consequently, cross-team percentage differences reflect a combination of tool behaviour, dataset composition and evaluator interpretation and should not be read as direct performance differences. The complete historical file-level CT-A/CT-B assignment and case-level denominators were not preserved in the publication package. Where figures diverge markedly for the same tool, the divergence is therefore treated as a limitation rather than a comparative ranking.

3.4 Performance analysis

The final stage of the methodology consisted of analysing the outputs generated during testing, with the aim of assessing the strengths and weaknesses of each tool category. This stage measured whether tools produced correct or incorrect

results, considering also factors such as usability, interoperability, and sustainability. In this sense, the analysis sought to capture the multidimensional nature of tool performance, reflecting both technical accuracy and the practical realities of investigative work.[18,21]

Performance was evaluated against a set of predefined criteria, treated as key performance indicators (KPIs). Accuracy was examined by comparing the tool's outputs with the expected results embedded in the synthetic datasets. For instance, in the case of manipulated images, accuracy referred to whether a tool successfully detected the alteration or misclassified the visual as authentic. Usability was assessed through the experience of the law enforcement testers, who recorded their observations on interface clarity, ease of operation, and the learning curve required for effective use. Interoperability referred to the extent to which outputs could be exported or integrated into broader analytical environments, such as open-source threat intelligence platforms. Coverage measured the alignment of each tool with specific adversarial behaviours drawn from frameworks like AMITT and DISARM, thereby assessing not just functionality but relevance to structured TTP detection. Finally, sustainability reflected whether tools were regularly updated, maintained active support communities, or showed signs of obsolescence, such as outdated interfaces or broken data pipelines.[22,24,25]

The available evaluation documentation uses a five-point ordinal scale (1-5) across seven dimensions: accuracy, coverage, speed/performance, ease of use, community feedback, customisability, and support/documentation. The five narrative indicators reported in this article relate to those dimensions as follows: usability corresponds primarily to ease of use; sustainability draws on community feedback and support/documentation; coverage and accuracy are reported directly; and interoperability was assessed qualitatively from export and integration behaviour. The retained records do not document a consistently administered System Usability Scale, inter-rater agreement test, or complete case-level output table for every tool. The data package provides the available input files, metrics documentation and explicitly encoded labels, but it does not permit independent reconstruction of every denominator or reported percentage. The percentages should therefore be interpreted as historical descriptive results rather than inferential estimates or fully reproducible benchmark scores.[35]

The evaluation process was iterative, involving both quantitative and qualitative elements. Quantitative measures were derived from the detection results themselves, such as success rates in identifying falsified content or accuracy in classifying suspicious domains. Qualitative insights were drawn from the logs maintained by the evaluators, who noted recurring errors, inconsistencies or usability issues. This combined approach is consistent with recent OSINT methodology studies, which emphasise that actionable intelligence emerges from combining measurable accuracy with user-centered evaluation.[20,23] It provided a broader picture of performance, while acknowledging that even tools with high detection accuracy may still fall short if their outputs are opaque or difficult to operationalise.

A further dimension of the analysis involved cross-comparison across categories. Since each category could be targeting different phases of FIMI operations, the analysis highlighted complementarities as well as gaps. For example, fact-checking tools proved strong at identifying recycled narratives but offered little support for technical attribution, while infrastructure verification tools contributed to attribution but could not address content falsification. Similar observations have been made in the literature, where the integration of heterogeneous OSINT tools is seen as key to offsetting categorical weaknesses.[18,22] The recognition of such complementarities was central to the construction of an optimised suite of tools, where individual strengths could be combined to offset categorical weaknesses.

By structuring the analysis in this way, the methodology ensured that findings were both descriptive and actionable. The performance assessment thus served as the bridge between the tool descriptions presented in Section 4 and the evaluative insights consolidated in Section 5, providing a systematic basis for determining which tools could be prioritised in the RESONANT suite and how the methodology could be applied to future detection instruments.

3.5 Ethical and legal safeguards

Testing activities were conducted under the governance of the project's DPIA and with attention to GDPR principles, proportionality and purpose limitation. The evaluation used project-prepared examples together with public and third-party reference material rather than live operational investigations. Some retained source files contain publicly available identifiers, including names, telephone numbers, addresses or email-header information; public availability does not remove the need for appropriate handling. The README identifies these categories and directs users to the original source terms and applicable data-protection requirements. Ethics-by-design considerations, including transparency, explainability and potential bias, informed the study but were not fully operationalised as measured performance indicators.

In the present study, ethics-by-design functioned as a guiding orientation rather than a measured outcome: no formal bias, transparency, or explainability metric was computed for the AI-based tools evaluated, and these criteria were not among the key performance indicators defined in Section 3.4. The epistemic implications of this gap are discussed in Section 5.3, and the operationalisation of transparency, explainability, and bias assessment as explicit indicators is identified as a priority for future iterations of the framework.

4. Results – Tool descriptions and mapping

The state-of-the-art review conducted in RESONANT identified a diverse set of tools relevant to the detection of TTPs linked to FIMI. To ensure analytical clarity, tools were grouped into functional categories, each corresponding to specific detection needs within frameworks such as AMITT and DISARM. The categorisation does not imply exclusivity, as many tools provide overlapping functionalities, but it highlights their principal contributions to different phases of adversarial operations. These categories include fact-checking tools, website credibility assessment tools, bot detection tools, image and video analysis tools, geo-IP verification tools, phone-related tools, social media analysis tools, and email verification tools. The following subsections provide descriptions of each category, including the core functionalities of representative tools and their relevance to TTP detection.

4.1 Fact-checking tools

Fact-checking tools represent the first line of defence in countering disinformation and identifying the recycling of debunked claims, a common tactic in FIMI campaigns. These tools allow investigators to cross-reference contested information against repositories of verified claims, thereby distinguishing between authentic and manipulative content. Their role extends beyond validating individual statements; they also enable the identification of recurrent adversarial procedures, such as the strategic reintroduction of false narratives into the information environment. This aligns with AMITT phases related to content creation and seeding, where adversaries exploit old or partially true claims to reinforce manipulative campaigns.

The utility of fact-checking tools lies in their accessibility and their ability to provide a rapid triage mechanism, aggregating claim reviews from multiple certified fact-checking organisations, enabling cross-validation of suspicious narratives, maintaining archives of debunked stories that often resurface during electoral cycles or crises or extending fact-checking into multimedia domains by combining keyword search with basic forensic analysis.

From a TTP detection perspective, fact-checking tools support the early identification of adversarial content falsification and repurposing techniques. They are particularly relevant for distinguishing between deliberate manipulative activity and genuine errors in reporting, an important distinction for law enforcement and policymakers. Nevertheless, RESONANT highlighted limitations: reliance on existing fact-checks can lead to gaps when novel narratives emerge, and the uneven geographical coverage of fact-checking initiatives leaves blind spots in certain languages and regions. Despite these constraints, fact-checking tools remain indispensable as an entry point into the layered detection process, feeding subsequent stages of investigation that rely on multimedia forensics, social media analytics, and threat intelligence platforms.

This constraint is not merely a matter of database coverage or tool maintenance; it reflects a structural property of the entire content-level detection layer. Fact-checking and credibility-assessment tools are content-dependent: they operate by matching a claim, domain, or artefact against previously encountered records or indexed knowledge. For genuinely novel content (emerging narratives, information voids surrounding breaking crises, or non-anglophone campaigns) such tools have, by construction, no prior reference against which to adjudicate, and their effective accuracy may degrade sharply. This distinguishes content-dependent detection from content-agnostic approaches, which target the structural and behavioural regularities of manipulation rather than the truth-value of specific content.[31] The maturity attributed to content-verification tools in Section 5.1 should therefore be read as conditional on the tested content resembling previously documented cases.

4.2 Website credibility assessment tools

Website credibility assessment tools extend verification practices from individual claims to the evaluation of entire online domains. Rather than focusing on the truthfulness of a single piece of content, these tools analyse and rate the broader information environment in which narratives circulate. RESONANT highlighted several representative tools, which provide structured evaluations of websites on parameters such as transparency of ownership, sourcing practices, adherence to journalistic standards and history of publishing disinformation.

The value of these tools for FIMI detection lies in their ability to flag low-credibility outlets that frequently serve as vectors for adversarial narratives. In many campaigns, content amplification relies on networks of fringe or pseudo-news

websites that are designed to mimic legitimate media outlets while promoting manipulative or misleading information. Credibility tools provide quick assessments of trustworthiness to help analysts identify such outlets at the early stages of an investigation and to prioritise them for deeper scrutiny using additional methods.

In the context of TTPs, these tools support the detection of adversarial behaviours such as the exploitation of unverified sources, establishment of false fronts, and use of proxy outlets for amplification. Their role is therefore complementary to fact-checking tools: while fact-checkers verify the accuracy of individual claims, credibility assessment tools address the structural dimension of disinformation ecosystems, highlighting domains that systematically generate or disseminate manipulative content.

4.3 Bot detection tools

Bot detection tools address one of the most prominent techniques used in FIMI: the deployment of automated or semi-automated accounts to amplify narratives and simulate grassroots activity. RESONANT reviewed tools that apply behavioural analysis and machine learning techniques to distinguish between authentic user activity and automated or AI-generated content.

These tools typically operate by analysing features such as posting frequency, network connections, content similarity, and linguistic markers. By detecting anomalies in these patterns, they can identify accounts that contribute disproportionately to the visibility of specific narratives. In some cases, bot detection platforms also provide dashboards and scoring systems that allow investigators to monitor the scale and distribution of suspected inauthentic activity.

The relevance of bot detection tools to TTPs lies in their ability to expose adversarial behaviours such as coordinated amplification, astroturfing, and the artificial inflation of engagement metrics. By doing so, they help analysts identify when narratives are being promoted not through organic user interest but through orchestrated manipulation.

However, as noted in RESONANT, these tools often face limitations imposed by platform restrictions, particularly as social media companies tighten access to their data through application programming interfaces (APIs). While they provide valuable indicators, their effectiveness is dependent on the quality and availability of underlying data. Nonetheless, bot detection remains a crucial element in any toolkit designed to uncover the amplification phase of FIMI campaigns.

4.4 Image and video analysis tools

Image and video analysis tools constitute a broad category within the RESONANT review, reflecting the centrality of multimedia content in FIMI. Visual material is frequently used to anchor manipulative narratives, either by presenting fabricated evidence, altering authentic media, or repurposing content out of its original context. Tools in this category were therefore subdivided into four functional groups: image search, AI-generated image detection, image alteration detection, and optical character recognition (OCR).

Image search tools enable investigators to determine whether a visual asset has appeared elsewhere on the internet, thereby helping to establish provenance and detect reuse of content in misleading contexts. Such tools allow analysts to trace images back to their earliest known appearance, which is particularly relevant for identifying TTPs associated with content repurposing and strategic recycling of imagery.

AI-generated image detection tools respond to the increasing prevalence of synthetic media, such as deepfakes or artificially generated photographs. This tool category was reviewed in RESONANT for the tools' ability to identify visual artefacts or statistical signatures typical of AI-generated content. Their role in TTP detection is linked to identifying fabrication techniques within the AMITT content creation phase, where adversaries employ synthetic visuals to enhance the credibility of disinformation campaigns.

Image alteration detection tools focus on identifying manipulations in authentic visual material, such as cropping, splicing, or retouching, and are capable of highlighting inconsistencies in compression, lighting, or metadata that may signal tampering. These tools help uncover TTPs associated with falsification and deceptive framing, where adversaries distort genuine evidence to support manipulative claims.

Optical character recognition tools extract embedded text from images or screenshots, enabling further analysis of content that would otherwise remain inaccessible to keyword searches. By converting visual text into machine-readable form, these tools facilitate both verification and content analysis, particularly when adversaries disseminate text-based disinformation in image formats to evade automated detection.

Taken together, image and video analysis tools form a vital component of the detection ecosystem. They provide investigators with the ability to interrogate visual material across multiple dimensions (provenance, authenticity, fabrication, and embedded content) thereby addressing adversarial procedures that exploit the persuasive power of multimedia.

4.5 Geo-IP verification tools

Geo-IP verification tools were included in RESONANT as a distinct category due to their role in linking digital traces to geographic locations. These tools provide analysts with information about the likely origin of an Internet Protocol (IP) address, often specifying country, region, city, and associated internet service provider. They enable investigators to conduct reverse lookups and determine the probable source of online activity.

The relevance of these tools for FIMI detection lies in their ability to connect online behaviours with physical infrastructure. Adversaries frequently rely on obfuscation techniques such as virtual private networks (VPNs) or proxy servers to conceal their location, yet patterns of IP use can still reveal clusters of activity or associations with particular geographic regions. Identifying these patterns contributes to the detection of adversarial TTPs, such as the use of anonymisation infrastructure or the coordination of campaigns across specific hubs.

From the perspective of technical attribution, Geo-IP tools provide supporting evidence that can be cross-referenced with other indicators. For example, repeated association of suspicious accounts with hosting infrastructure in particular regions may reinforce assessments of coordinated operations. Although such findings must be treated with caution, given the prevalence of obfuscation, they remain valuable for situating FIMI activity within a broader operational context. Thus, Geo-IP verification tools complement content- and network-focused detection methods by adding a geographic dimension to the analysis.

4.6 Phone-related tools

Phone-related tools were assessed in RESONANT as part of the broader detection ecosystem due to the recurring use of telephone numbers in disinformation campaigns and fraudulent online activity. Such numbers may be embedded in websites, social media posts, or messaging platforms as supposed points of contact, often intended to lend credibility to manipulative narratives or to facilitate direct engagement with targeted audiences.

These tools provide services such as carrier identification, reverse lookups and risk assessment. They can reveal whether a number is genuine, spoofed, or disposable and in some cases, offer insights into geographic origin or associations with known spam and fraud databases.

The contribution of phone-related tools to FIMI detection is relatively specialised but nonetheless, significant. They support the identification of adversarial TTPs linked to credibility-building through false contact information, fraudulent engagement tactics, and the use of disposable communication channels. Investigators can determine whether contact details are part of a broader manipulation strategy by verifying or discrediting phone numbers, thereby complementing other verification and attribution tools. While their relevance is narrower compared to categories such as social media analytics or multimedia forensics, phone-related tools address a specific detection need and contribute to building a comprehensive picture of adversarial operations. In doing so, they add another layer of evidence that can strengthen technical attribution when combined with other open-source indicators.

4.7 Social media analysis tools

Social media analysis tools were included in RESONANT as a distinct category due to the centrality of platforms such as Facebook, X (formerly Twitter), and Reddit in FIMI operations. These tools are designed to extract, filter, and monitor content from specific platforms, thereby providing insights into narrative dissemination, engagement dynamics, and potential signs of coordinated inauthentic behaviour.

From a TTP perspective, such tools are particularly relevant for detecting amplification and coordination techniques, since social media remains the primary arena for the spread of manipulative narratives. However, RESONANT also documented important limitations: many of the reviewed tools either returned incomplete or false results or were no longer fully functional due to platform restrictions on data access. This situation highlights a structural vulnerability in FIMI detection, where analysts are dependent on tools whose effectiveness can quickly diminish as platforms adjust their policies and technical environments. Despite these constraints, social media analysis tools remain essential for monitoring the real-time evolution of online campaigns, even if their long-term sustainability is uncertain.

4.8 Email verification tools

Email verification tools formed the final category examined in RESONANT. These tools are designed to validate the authenticity and deliverability of email addresses, detect the use of disposable or temporary domains, and in some cases perform reverse lookups to associate addresses with user profiles or organisations. The value of email verification tools for FIMI detection lies in their ability to expose adversarial TTPs that rely on fraudulent contact points or false identity markers. Manipulative campaigns often incorporate email addresses into websites, social media accounts, or outreach material as a means of lending credibility or facilitating direct interaction with targeted audiences. Investigators can quickly identify when such contact details are fabricated, high-risk or linked to disposable providers by verifying the status of these email addresses.

Similar to phone-related tools, their application is narrower compared to other categories such as multimedia analysis or social media monitoring. Email verification tools provide a complementary function within the overall detection process. They contribute to the broader objective of technical attribution by validating or discrediting contact vectors embedded within manipulative campaigns. In doing so, they add an additional evidentiary layer that, when combined with other open-source indicators, helps build a more comprehensive understanding of adversarial operations.

5. Results and analysis

The structured testing of OSINT and verification tools within the RESONANT framework generated descriptive findings concerning technical behaviour and operational usability. The evaluation was conducted collaboratively by cyberteams A (CT-A) and B (CT-B), consisting of Hellenic Police and Ukrainian Police experts, each working with distinct selections of the mixed-source evaluation corpus under their respective operational conditions. The results illustrate behaviour on the historical evaluation material but should not be treated as statistically representative estimates of general FIMI-detection performance. The observations reflect the specific testing period, available tool versions, dataset composition and evaluator context. OSINT tools evolve rapidly and their functionality, access conditions and performance may change because of software updates, API modifications or platform policies. The results are therefore indicative of methodological applicability rather than fixed performance rankings. The following subsections analyse tool effectiveness and limitations, propose an optimised suite of instruments, and reflect on practical challenges encountered during testing.

5.1 Tools' effectiveness and limitations

The first category examined was **fact-checking tools**, which play a fundamental role in the early triage of false or recycled narratives. Across both evaluations, results showed clear disparities in performance. GPT-4o produced the strongest scores within the historical fact-checking tasks, correctly identifying around 90 percent of true and 58 percent of false claims in the CT-A tests, and achieving 94 percent accuracy on true and 95 percent on false claims in the CT-B evaluation, with only minor false-positive rates. The Factual ranked second in the CT-A results, correctly classifying about 52-55 percent of true and 47 percent of false claims, though it showed higher uncertainty levels (up to 45 percent undefined). The CT-B results were broadly similar for true-claim detection, though other fact-checking repositories performed inconsistently on false claims. Traditional databases such as Snopes, Google Fact Check Explorer, PolitiFact, Truth or Fiction, AFP Fact Check, and Fake News Debunker provided limited utility, often returning undefined outputs or failing to match phrasing with their existing entries. These findings indicate that GPT-4o and The Factual were the most informative fact-checking tools on the historical, mixed-source evaluation material, while conventional repositories remain useful primarily for verifying widely circulating or previously indexed claims. They should not be read as evidence that GPT-4o is the most reliable operational fact-checking instrument for novel, emerging, or out-of-distribution FIMI content.

Four caveats qualify these figures. First, because parts of the evaluation corpus derive from or link to publicly documented disinformation cases and third-party collections, GPT-4o may achieve high apparent accuracy by recognising content encountered during training rather than by independently verifying it. The percentages should therefore not be interpreted as evidence of reliable fact-checking on novel or out-of-distribution FIMI content. Second, the exact historical GPT-4o snapshot, prompt, temperature and tool/browsing configuration were not retained, which limits model-level reproducibility. Third, the pronounced divergence between CT-A and CT-B (most notably GPT-4o's 58 percent versus 95 percent false-claim accuracy) cannot be attributed to the tool alone because the teams used distinct selections and operated under different conditions (Section 3.3). Fourth, no formal inter-rater reliability coefficient was computed and the complete case-level denominator record is unavailable. Accordingly, the accuracy figures should be read as historical agreement with the project-defined labels available at the time, not as externally validated or independently reproducible estimates of general performance. The deposited package improves transparency about the available input pool and metrics but does not make every denominator and output auditable.

The evaluation of **website-credibility assessment tools** further highlighted the uneven quality of existing resources. While the conceptual value of these tools lies in their ability to provide contextual judgments on the reliability of entire domains, their practical effectiveness was modest. NewsGuard, despite being widely cited in policy discussions, could not be tested due to its subscription-based model, and Schema.org's ClaimReview lacked the necessary functionality for systematic evaluation. The Know News browser extension demonstrated very limited accuracy, correctly identifying credible or non-credible domains in only about 16 percent of cases. By contrast, MyWOT (Web of Trust), which was integrated after the main testing phase, proved more useful in assessing overall domain reputation. These findings illustrate a broader limitation of credibility-assessment approaches: although they can help highlight recurrent low-credibility sources, their coverage remains narrow and they often require human interpretation to validate results.

When evaluating **bot detection tools**, the testing process encountered major structural limitations. BotSentinel, once a widely used platform, discontinued its API service and was therefore unavailable for testing. GPTZero, designed primarily for detecting AI-generated text, was tested but could not be effectively applied to identifying automated social-media accounts. The limitation stemmed from the requirement to manually extract posts and comments from platform profiles—an action restricted under Twitter's terms of service. These challenges highlight a broader issue in bot detection: effective analysis depends on access to authentic behavioral data, yet such access is increasingly limited by platform policies and API restrictions. Consequently, while bot detection remains an essential component of FIMI monitoring, current tools cannot be reliably deployed without sustainable and compliant data-access mechanisms.

The **image and video analysis** category showed strong and consistent performance, making it one of the most reliable layers of the toolkit. **Reverse-image-search tools**, including *Google Images*, *TinEye*, and *Search by Image*, achieved 100 percent accuracy, successfully identifying all instances of manipulated or recycled images within the dataset. Tools designed to detect AI-generated content produced more mixed results: *ContentAtScale.ai* and the *Umm-maybe* model hosted on Hugging Face correctly identified about 72 percent of real images and 52–56 percent of AI-generated ones, while *Illuminarty.ai* performed poorly, detecting only around 20 percent of fake images and generating a high number of false positives. **Image-alteration-detection tools**, such as *FotoForensics* and *29a.ch Photo Forensics*, also showed uneven but useful performance, both detected 100 percent of genuine images, while achieving 60 percent and 40 percent accuracy, respectively, on altered ones, yet their results required expert interpretation. **OCR** tools performed very strongly: *Google Lens* achieved 92 percent accuracy, *ChatGPT OCR* 88 percent, and *OCR.space* 68 percent, clearly outperforming older tools such as *NewOCR.com* (28 percent) and *OnlineOCR.net* (16 percent). Taken together, these results indicate that while multimedia forensics is not immune to errors, the current ecosystem of tools for reverse search, AI-image detection, alteration analysis, and text extraction is relatively mature and can be effectively operationalised for investigative use.

The **infrastructure-level tools** produced more varied results. **Geo-IP detection platforms** displayed a moderate degree of accuracy. *MaxMind GeoIP* failed to operate during testing, while *IPFingerprints* achieved only 36 percent accuracy in identifying the correct country. By contrast, *IPLocation.net* and *Thatsthem IP Lookup* reached about 70 percent accuracy, indicating that while IP geolocation cannot serve as a standalone attribution method, it provides a useful supporting layer for situating suspicious activity within broader geographic patterns. **Phone-number-verification tools** showed clearer contrasts: *WhoCalled.com* delivered the strongest results, correctly matching approximately 89 percent of tested numbers in both evaluations and, in several cases, even identifying partial owner information. *IPQualityScore's* validator achieved about 65 percent accuracy in the CT-A results but did not return usable data in the CT-B evaluation, as it required manual entry of country codes for each query. *Truecaller*, despite its popularity, lacked the necessary functionality in its free version and could not be properly tested without login credentials. Overall, the results confirm that *WhoCalled.com* offers the highest reliability for phone-number analysis, while other tools remain constrained by technical or access limitations.

Finally, **social-media-analysis tools** proved to be the least reliable category. Many of the platforms previously used for extracting or monitoring social-media data, such as *CrowdTangle*, *Ad Observer*, and *SearchIsBack*, are now deprecated, restricted, or entirely non-functional. Tools designed for *Twitter*, including *Advanced Search* and *Geo-Search*, often returned incomplete or false results, while *Facebook*-based searches and numerous *Reddit* utilities such as *Subreddit-Stats*, *Reddit User Analyser*, and *Apify Reddit Scraper* either failed to respond or produced accuracy levels below 40 percent. Several others, including *Redective*, *Keyworddit*, and *SowSearch*, produced errors or unreliable outputs. Overall, the category was marked by discontinued services, false data, and API restrictions imposed by major platforms. This reflects a broader structural issue: as social networks tighten data access, third-party analytical tools rapidly become obsolete. Consequently, while social-media analysis remains essential for FIMI detection, the current landscape is unstable and requires renewed development of sustainable, API-driven, and institutionally supported solutions.

The evaluation of **email-verification tools** produced mixed but informative results. Among the tools tested, *Mailbox-Validator* and *Thatsthem Reverse Email Lookup* delivered the most consistent performance, correctly identifying around

21–22 percent of valid addresses and providing detailed analytical parameters. *EmailHippo* produced useful outputs but showed a higher rate of unverifiable results (18 percent) and lower overall accuracy. *Emailable* and *EmailValidator.com* could not be tested effectively due to account or subscription restrictions. In the CT-B tests, *MailboxValidator* again demonstrated stable functionality, while *Thatsthem* provided only partial informational outputs without explicit verification results. Overall, the findings indicate that while several email-verification tools remain limited by access or functionality constraints, *MailboxValidator* and *Thatsthem* together offer the most reliable combination for validating contact information within OSINT investigations.

Integrated and cross-domain evaluations

Beyond single-function testing, a series of integrated evaluations was conducted to measure interoperability and cross-domain performance. In the **Comprehensive Media Analysis**, the CT-A achieved strong, domain-specific results across all stages: *29a.ch Photo Forensics* reached 100 percent accuracy in image and video analysis, *Originality.ai* achieved 89–90 percent accuracy in content verification, and *EmailHippo* succeeded in all contact-validation tests. CT-B obtained comparable but more qualitative outcomes: *29a.ch Photo Forensics* effectively detected image alterations such as cloning but provided limited contextual detail; *Truecaller* functioned well as a mobile application for identifying spam or malicious phone numbers, and *The Factual* produced a "Moderately Credible" score of 51 percent for article reliability, performing effectively on most standard-length articles. Taken together, these findings demonstrate that well-configured combinations of content, media, and contact-verification tools can deliver near-complete analytical coverage, though interpretation still requires a degree of manual expertise.

The **Disinformation Campaign Analysis** revealed mixed levels of effectiveness across the tested tools. For CT-A, *The Factual* proved highly reliable, achieving 100 percent success in article analysis, while *Search by Image* reached 43 percent accuracy in image and video verification, and *Know News* scored 0 percent in social-interaction analysis. These results highlight the strong performance of text-based verification tools but also the need to integrate more robust image-analysis solutions and alternative methods for assessing online interactions. CT-B obtained generally effective outcomes, particularly for article analysis, with *Know News* performing well in credibility assessment and *FotoForensics* offering detailed image insights that nevertheless required user expertise to interpret accurately. Overall, the findings confirm that while content-verification tools are mature and reliable, image forensics and social-interaction mapping remain dependent on operator skill and would benefit from more advanced, automated capabilities.

The **Cross-Platform Disinformation Analysis** produced mixed results. In CT-A's evaluation, *The Factual* recorded 0 percent success in article verification, while *MediaBiasFactCheck* achieved 100 percent accuracy in social-media analysis. *CrossRef Search* failed to generate relevant matches (0 percent), and *MonkeyLearn Sentiment Analysis* produced a 40 percent positive versus 60 percent negative sentiment balance. CT-B reported more stable outcomes, with *The Factual* performing effectively in assessing article reliability. Overall, the findings indicate that while textual and sentiment-analysis tools can yield valuable insights, the ability to establish reliable cross-platform correlations remains limited, highlighting the need for better-integrated analytical frameworks within the current OSINT ecosystem.

Finally, the **Multi-Source Disinformation Campaign Analysis** assessed the toolkit's capacity to detect and contextualize disinformation across complex geopolitical narratives. In CT-A's assessment, *The Factual* achieved 42 percent success in article analysis, *Scribbr AI Detector* reached 14 percent in social-media analysis, and *X.com Explore* delivered 100 percent accuracy in cross-reference detection. CT-B did not report available data for this phase. Overall, the results indicate that while content-verification tools can provide partial success in large-scale misinformation environments, automation and cross-modal integration remain limited, underscoring the continued need for human oversight and contextual reasoning in comprehensive disinformation analysis.

OSINT tools: Limitations

Although OSINT tools often seem to be very powerful while conducting digital investigations, due to the automating and fast way of functioning, still they present some significant limitations.

One major limitation is their volatility. In general, digital platforms, social networks, and online repositories constantly change their structures, restrict access, or modify interfaces, depending on various administrative, legal and other reasons. Of course, this also has an impact on an OSINT tool, as these changes may most likely prevent it from functioning. Thus, this constantly unstable situation should be taken into consideration by all interested persons/organizations, who should not rely heavily on any single automated solution.

Next, another limitation is the narrow scope of data coverage. For example, an OSINT tool may be designed to focus on a specific category, such as IP addresses, domain records, usernames, or social media accounts, while leaving many others uncovered, of high importance. This can lead to the conclusion that no single tool can ever provide a complete intelligence picture, and important information can easily be missed if investigations rely solely on tool-generated results.

Another limitation of OSINT tools is the reliability of the generated data. Many tools rely on old, outdated or limited/incomplete sources, which often results in misleading and inaccurate findings. For this reason, it is vital to properly validate and cross-check all resulting information prior to its use or dissemination.

Furthermore, a common risk that someone can fall into is the lack of security measures. This risk emerges from the fact that many OSINT tools interact directly with target platforms (domains), leaving residual evidence in the target log files. This could potentially raise an alert for the administrator, indicating that an investigation is underway. For this reason, proper operational security practices, such as anonymisation, virtual environments, and sock puppets are essential to avoid these risks.

The previous point is reinforced by the fact that there is often an "automation mentality" that gives the impression that OSINT tools are simple, requiring only pressing a button and wait for the results. This creates a false sense of security as, for sure, the tools can assist in the intelligence gathering process, but they cannot replace human creativity, analytical thinking, and reasoning.

Given the rapid evolution of OSINT platforms and access conditions, repeated testing over time would likely produce different operational outcomes.

In summary, OSINT tools are valuable for increasing speed and efficiency, but they have clear limitations. They cannot ensure complete coverage, absolute accuracy, or replace human judgment. Successful OSINT work relies on the investigator's skills, operational security, and solid methodology, with tools acting as supporting resources rather than definitive solutions.

5.2 Optimised tool suite

On the basis of the comparative evaluation, an optimised suite of tools was proposed for integration into the RESONANT detection environment. At the content-verification level, GPT-4o and The Factual were retained as the most accurate and scalable fact-checking instruments, with traditional repositories maintained in reserve for redundancy and historical coverage. Originality.ai was included as a complementary resource for automated content verification. For website credibility, MyWOT was prioritized as the most consistently usable tool, with Know News included as a secondary option. In the domain of multimedia forensics, Google Images, TinEye, and Search by Image were selected for reverse search; ContentAtScale.ai and the Umm-maybe model (Hugging Face) were retained for AI-image detection; FotoForensics and 29a.ch Photo Forensics were integrated for alteration analysis; and Google Lens and ChatGPT OCR were identified as the strongest performers for text extraction. At the infrastructure level, IPLocation.net and Thatsthem IP Lookup were recommended for IP geolocation, WhoCalled.com for phone verification, and a combination of Mailbox-Validator and Thatsthem Reverse Email Lookup for email verification.

Although bot detection and social-media analytics remain structurally important, the evaluation concluded that current tools are either deprecated or insufficiently reliable, underscoring the need for further development and integration of sustainable, API-driven solutions. The same applies to cross-platform and multi-source disinformation analysis, where tool interoperability and automation are still limited.

This optimised suite reflects a layered and complementary approach to tool selection. No single tool or category was sufficient to address the full spectrum of adversarial tactics, techniques, and procedures. However, by combining high-performing instruments across content, platform, and infrastructure levels, it is possible to build a detection ecosystem that compensates for individual weaknesses. The proposed suite is therefore not a static configuration, but a dynamic baseline, designed to evolve as tools are updated, discontinued, or replaced.

The recommendation to retain GPT-4o as a content-verification triage instrument must be qualified by the documented tendency of large language models to exhibit systematic biases in evaluative tasks. These include source-framing effects, in which the stated origin of a text (independent of its content) shifts the model's credibility judgement, as well as broader ideological and political biases in language-model evaluation systems.[32,33] Because such biases arise from regularities in training data and alignment processes, they cannot be reliably removed through prompting alone, and they are particularly consequential in geopolitically sensitive FIMI contexts where attribution-adjacent judgements are at stake. GPT-4o is

also among the least transparent instruments in the proposed suite. Its inclusion is therefore conditional on operational safeguards: content should, where feasible, be presented to the model without source attribution, and its outputs should be cross-validated against source-agnostic or structured-evidence tools rather than relied upon as standalone verdicts.

Because the suite combines tools that produce structured evidence for analyst interpretation with tools that produce direct verdicts (such as fact-check classifications and credibility scores), its operational use is governed by a minimal decision protocol. First, verdict-producing outputs are treated as investigative leads rather than conclusions and must not, on their own, ground an attributional or evidentiary claim. Second, any verdict intended to inform an investigative action requires confirmation from at least one independent, and preferably source-agnostic, source before use. Third, human analyst review takes precedence over automated output wherever the two diverge, consistent with the principle that tools support but do not replace analytical reasoning (Section 5.1). This protocol distinguishes the proposed suite from a mere aggregation of high-performing tools and treats the automation-mentality caution as a structural safeguard rather than a peripheral note.

Finally, because several of the recommended instruments are commercial, proprietary, subscription-based, or dependent on platform access, and because tool functionality changes rapidly, this suite represents a time-bound baseline valid for the tested period and tool versions rather than a set of rankings generalisable beyond the evaluation window.

5.3 Practical challenges

In addition to technical performance, the evaluation process exposed several practical challenges that must be considered in the operationalisation of OSINT methodologies. The first challenge relates to **sustainability**: many tools tested were found to be discontinued, inaccessible, or locked behind subscription models, sometimes in the middle of the evaluation itself. This instability undermines the reproducibility of results and highlights the need for a more institutionalized approach to tool provision, potentially through public-private partnerships or open-source development.

A second challenge concerns **interoperability**. While the testing aligned tool outputs with structured frameworks such as DISARM and AMITT, few tools natively export results in machine-readable or standardized formats. This lack of integration increases the burden on analysts and limits the scalability of detection efforts. Without interoperability, it is difficult to link outputs across categories or feed them into broader intelligence platforms.

Third, the evaluation revealed persistent **coverage gaps**, particularly regarding multilingual detection and non-Western information environments. Most fact-checking repositories are heavily biased towards English-language content, while many credibility tools assess primarily Western news outlets. This creates blind spots that adversaries can exploit by targeting under-represented linguistic or regional contexts.

Significant ethical and legal constraints arise in the use of OSINT tools, particularly concerning bot detection and the scraping of data from social media platforms. These challenges are especially relevant in operational settings, where strict compliance with data protection regulations (such as the GDPR) and adherence to platform terms of service are required. While controlled, non-operational datasets can help mitigate some of these risks during development and testing, they do not eliminate the legal and ethical complexities involved in real-world applications. As emphasised in the attribution literature, technical observables identified through OSINT tools can support investigations but cannot, on their own, establish responsibility for information influence operations, which ultimately requires contextual, political, and legal assessment.[34] Operational usability is another critical factor that can affect the effectiveness of OSINT tools. Even when tools are technically accurate, their practical value may be reduced by unclear user interfaces, non-transparent scoring systems, or steep learning curves. These issues can limit accessibility and hinder adoption by investigators and analysts. Beyond legal compliance, ethical considerations play a vital role in the responsible use of OSINT. Analysts must ensure that their actions are necessary, proportionate and respectful of fundamental rights. This includes avoiding techniques that could be seen as intrusive, manipulative, or misleading. In the absence of universally accepted international standards, OSINT practitioners must rely on sound professional judgment and uphold principles of transparency, accountability and integrity in their work.

A distinction should be drawn between legal-compliance ethics and epistemic ethics. The safeguards described in Section 3.5 (GDPR compliance, proportionality, data minimisation, and the project DPIA) address the lawfulness of data processing. They do not, however, address a separate ethical question raised by the deployment of verdict-producing algorithmic tools in investigative contexts: when an opaque model such as GPT-4o classifies a claim as false and that output informs a law enforcement judgement, the tool functions as an epistemic authority whose reasoning is not, in principle, open to scrutiny or contestation. Non-transparency in verdict-producing tools is therefore not merely a usability concern but an epistemic-ethics concern. Accordingly, the operational protocol set out in Section 5.2 restricts the weight

that opaque, verdict-type outputs may carry and requires that they be corroborated and remain subject to human override before informing any investigative action.

Taken together, these challenges highlight the need for a robust, flexible, and ethically grounded framework for OSINT testing and validation. The RESONANT methodology provides an initial foundation by consolidating tool evaluations and aligning them with structured frameworks. However, further efforts will be required to ensure sustainability, interoperability, and accountability in the rapidly evolving ecosystem of FIMI detection tools.

6. Conclusion and future research

This article has presented a structured methodological framework for testing OSINT tools in the detection of FIMI-linked operations. Building on established taxonomies such as DISARM, AMITT, and MITRE ATT&CK, the framework consolidates conceptual, operational, and evaluative elements into a coherent process. Its primary contribution is an empirically and operationally informed workflow for selecting, applying and interpreting heterogeneous OSINT tools. The framework is transferable at the level of its staged procedure and evaluation dimensions; exact replication of all reported percentages is limited by the incomplete historical run record. It therefore advances methodological transparency without claiming a fully reproducible benchmark or formal systematic review.

A key contribution is the articulation of a multi-layered detection logic, distinguishing content-level verification, platform-level activity analysis and infrastructure-level attribution. Previous research highlights the importance of these layers, yet assessments often remain fragmented, focusing on individual tools or isolated use cases. The RESONANT methodology operationalises this layered understanding, demonstrating how complementary strengths across categories can mitigate inherent weaknesses. This offers researchers and law-enforcement agencies a more coherent lens through which to interpret the ecosystem of FIMI detection tools.

The study also illustrates the practical value and limitations of a controlled, mixed-source evaluation corpus. Combining project-prepared examples with selected public and third-party reference material can support diverse analytical tasks, but it also creates provenance, licensing, privacy, contamination and representativeness constraints. Publication of the available input corpus, metrics documentation and README improves transparency about what was retained. It does not constitute a complete replication package because the full CT-A/CT-B mapping, case-level outputs, denominators, logs and model configurations are unavailable. This boundary is material to the interpretation of the findings and to future implementations of the framework.

At the same time, the study highlights several limitations that inform future work. First, the controlled mixed-source corpus cannot fully capture the dynamic unpredictability of evolving information manipulation tactics, and its predominantly English material does not support multilingual performance claims. Second, incomplete preservation of the historical CT-A/CT-B case mapping, run-level outputs, denominators and exact model configurations prevents exact replication of every reported result. Future evaluations should preserve these artefacts prospectively through stable case identifiers and versioned execution logs. Third, OSINT tools operate within a rapidly changing digital landscape in which API restrictions, platform policies and data availability affect performance over time. Fourth, the methodology focuses on detection rather than classification of intent or attribution, which require further analytical and legal input.

Looking forward, the proposed framework can support a number of research and operational pathways. Future studies could expand the testing protocol to integrate machine-learning-driven tools, automated pipeline orchestration and cross-platform detection capabilities. There is also scope to explore interoperability with threat-intelligence platforms and harmonisation with EU-level standards for analysing information manipulation activities. Finally, applying the methodology across additional languages, geopolitical contexts, and sociotechnical environments would enhance its generalisability and support comparative research across the FIMI domain.

Beyond extending the existing paradigm, future work must confront a structural constraint that the present results make visible. The two detection categories most dependent on platform data (bot detection and social media analytics) were rendered largely non-functional by platform API restrictions, and this cannot be resolved within a third-party tooling model. Addressing it raises a question that is regulatory rather than technical, namely whether effective FIMI detection at scale requires public-interest or institutionally guaranteed access to platform data. This is identified here as the most pressing open problem for the field. Equally, the recognition that tools support but cannot replace human analytical judgement should be treated as a design principle for any future detection suite rather than as a concluding caveat.

In sum, this study contributes to strengthening the evidence base for OSINT tool evaluation in the context of FIMI detection. By offering a reproducible, ethically grounded, and operationally informed framework, it advances both

scholarly understanding and practical capabilities for countering information manipulation, supporting broader EU efforts to enhance democratic resilience and information security.

Data availability – Grant information

Zenodo. *Detecting FIMI: A methodological framework for testing OSINT tools in TTPs detection.* https://doi.org/10.5281/zenodo.19045753.[35]

The study used a controlled, mixed-source evaluation corpus assembled for the Horizon Europe project RESONANT (Grant Agreement No. 101132439). The retained archive combines project-prepared illustrative material, labelled text and bot-detection data, public website and infrastructure reference data, historical email and phone-related inputs, and links to third-party multimedia datasets and documented disinformation cases. The textual material is predominantly English. Some retained source files contain publicly available names, telephone numbers, addresses, email-header information or other identifiers; users must therefore apply the relevant source terms, data-protection requirements and purpose limitations.

The available data package has been deposited in Zenodo (link above and at [32]). It contains the retained evaluation input pool, the metrics and documentation files, and a README describing the archive. Principal components include 12,228 fake-labelled and 9,762 real-labelled non-empty text items; reduced subsets of 32 and 42 items; 50,000 bot-detection records (25,018 label 1 and 24,982 label 0); 35 website records; 100 email files; two phone-related tables containing 441 and 100 records; three five-record social-media examples; IP/ASN reference data and samples; external image-dataset links; and four link-based fact-checking scenarios. Ground truth is available only where explicitly encoded by fake/real file membership, the Bot Label field, or FACT/FALSE scenario notation. The archive does not preserve a uniform difficulty score or TTP label for every item, a complete case-level CT-A/CT-B mapping, all tool outputs and logs, the exact denominator behind every reported percentage, or the complete GPT-4o configuration. It therefore supports inspection of the available materials and transfer of the methodological template, but not exact independent replication of every numerical result. Requests for any additional material should be addressed to the corresponding author at the Center for Security Studies (KEMEA) (k.margaros@kemea-research.gr) and considered under the project's Data Management Plan and applicable approvals.

Information on the OSINT tools considered in this study was collected through a structured multi-source approach drawing on publicly available materials, academic and grey literature, EU-funded project outputs, repositories such as AMITT, and the operational expertise of RESONANT partners. The OSINT tools themselves are publicly accessible, and their access points are documented in RESONANT Deliverable D2.1: State of the Art of Tools for TTPs Detection and Analysis.

The archive includes public and third-party source material. Source links are provided where retained, but provenance and redistribution terms were not preserved uniformly for every historical file. Absence of a source entry must not be interpreted as RESONANT ownership. Users are responsible for complying with original licences, database rights and terms of use. The README documents known provenance, privacy and data-quality limitations. Project handling was governed by the RESONANT DPIA; the public availability of source information does not remove the need for proportionate and lawful processing.

Ethics and consent

Ethical approval and consent were not required for this study. The research did not involve human participants or live operational investigations. The evaluation used project-prepared examples together with public and third-party reference material; some retained files contain publicly available identifiers, handled under the project DPIA in line with GDPR proportionality and purpose-limitation principles.

Acknowledgements

The author would like to thank all partners of the Horizon Europe project RESONANT (Grant Agreement No. 101132439) for their collaboration and contributions to the methodological development described in this article. Special appreciation is extended to the Hellenic Police (HP) and the Kharkiv National University of Internal Affairs/ Ministry of Internal Affairs of Ukraine (KhNUIA/MoIA) for their operational insights and participation in the testing activities. The author also acknowledges the support of the project coordinator Fondazione per la Ricerca sulla Migrazione e Integrazione delle Tecnologie (FORMIT) for facilitating the project's overall implementation and scientific integration.

Finally, the author wishes to thank colleagues at the Center for Security Studies (KEMEA) for their valuable feedback during the preparation of this manuscript.

References

1. EEAS: *2022 report on EEAS activities to counter FIMI.* European External Action Service; 2023.
Reference Source
2. EEAS: *How to Detect & Analyse Identity-Based Disinformation/FIMI: A Practical Guide to Conduct Open Source Investigations.* European External Action Service; 2024.
Reference Source
3. Government of Canada: *Foreign information manipulation and interference targeting Canada's democratic processes: Special report for 2019–2023.* Canadian Security Intelligence Service (CSIS) & Communications Security Establishment (CSE); 2023.
4. CogSec Collaborative, & Threet LLC: **AMITT TTP guide.** n.d.
Reference Source
5. Dou W, Wang X, Ribarsky W, *et al.*: *Event detection in social media data.* University of North Carolina at Charlotte & IBM Almaden Research Center; 2012.
6. Chinnov A, Kerschke P, Meske C, *et al.*: **An overview of topic discovery in Twitter communication through social media analytics. In Proceedings of the Twenty-first Americas Conference on Information Systems (AMCIS 2015).** 2015.
7. Ferrara E, Varol O, Davis C, *et al.*: **The rise of social bots.** *Commun. ACM.* 2016; **59**(7): 96–104.
Publisher Full Text
8. Demetzou K: **Data protection impact assessment: A tool for accountability.** *Computer Law & Security Review.* 2019; **35**(6): 105339.
Publisher Full Text
9. Sabir MI: **Assessing attribution and credible deterrence in cyberspace.** *J. Cyber Policy.* 2024; **9**(1): 44–63.
Publisher Full Text
10. Tsagourias N, Farrell C: **Cyber attribution: Technical and legal approaches and challenges.** *Research handbook on international law and cyberspace.* Edward Elgar Publishing; 2020; pp. 212–229.
Publisher Full Text
11. CogSec Collaborative, & Threet LLC: **AMITT user guide.** n.d.
Reference Source
12. CogSec Collaborative, & Threet LLC: **AMITT incident list.** n.d.
Reference Source
13. Rani N, Bansal A, Batra S, *et al.*: **A comprehensive survey on advanced persistent threat attribution with enabling paradigms and challenges.** *J. Netw. Comput. Appl.* 2024; **239**: 103738.
Publisher Full Text
14. Aral S, Dellarocas C, Godes D: **Social media and business transformation: A framework for research.** *Inf. Syst. Res.* 2013; **24**(1): 3–13.
Publisher Full Text
15. Kane GC, Alavi M, Labianca GJ, *et al.*: **What's different about social media networks? A framework and research agenda.** *MIS Q.* 2014; **38**(1): 275–303.
Publisher Full Text
16. Qin Z, Zhou Y, Wu T, *et al.*: **How rumors spread and stop over social media: A multi-layered communication model.** *HCIS.* 2015; **5**(1): 1–21.
Publisher Full Text
17. Mattingly D, Yao E: **How soft propaganda persuades.** *Comp. Pol. Stud.* 2022; **55**(14): 2427–2455.
Publisher Full Text
18. Dokman T, Ivanjko T: **Open source intelligence (OSINT): Issues and trends.** *INFuture 2019: Knowledge in the digital age.* University of Zagreb; 2019; pp. 191–196.
Publisher Full Text
19. Hwang J, Kim S, Lee J: **Current status and security trend of OSINT.** *Journal of Information and Security.* 2022; **2022**(4): 345–360.
Publisher Full Text
20. Tuominen S: **OSINT: Intelligence and its application.** *Poliisiammattikorkeakoulun raportteja.* 2019; **138**: 1–74.
Reference Source
21. Glassman M, Kang M: **Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT).** *Comput. Hum. Behav.* 2011; **28**(2): 673–682.
Publisher Full Text
22. Nobili M: **A systematic review of OSINT tools for social engineering.** *J. Digit. Forensic Secur. Law.* 2023; **18**(1): 1–20.
Publisher Full Text
23. Julan CI, Togan M: **Methodologies for retrieving and processing information from open sources (OSINT).** *Journal of Military Technology.* 2023; **6**(1): 39–44.
Publisher Full Text
24. Hassan N, Hijazi R: *Open source intelligence methods and tools: A practical guide to online intelligence.* Apress; 2018.
Publisher Full Text
25. Meredith A: *The OSINT handbook: A practical guide to intelligence gathering.* Routledge; 2024.
26. Adetokunbo A, Awodele O, Ogbonna AC: **A review of the waterfall model and object-oriented approach.** *Int. J. Eng. Res. Appl.* 2013; **3**(6): 729–732.
27. Andrei B: **A study on using waterfall and agile methods in software project management.** *Journal of Engineering Studies and Research.* 2019; **25**(3): 106–111.
Publisher Full Text
28. Aroral H: **Waterfall process operations in the fast-paced world.** *International Journal of Creative Research Thoughts.* 2021; **9**(6): 2324–2330.
29. Senarath U: **Waterfall methodology, prototyping and agile development: A comparative study.** *Int. J. Adv. Res. Comput. Sci.* 2021; **12**(3): 1–7.
Publisher Full Text
30. Eason O: **An analysis of waterfall to agile methodology.** *International Journal of Engineering Research and General Science.* 2016; **4**(3): 781–788.
31. Redaelli S, Spitale G, Germani F: **The structural fingerprints of disinformation: A content-agnostic framework.** *Ethics Inf. Technol.* 2026; **28**: 29.
Publisher Full Text
32. Germani F, Spitale G: **Source framing triggers systematic bias in large language models.** *Sci. Adv.* 2025; **11**(45): eadz2924.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
33. Fulay S, Brannon W, Mohanty S, *et al.*: **On the relationship between truth and political bias in language models.** 2024; arXiv: 2409.05283.
Publisher Full Text
34. Pamment J, Smith V: *Attributing Information Influence Operations: Identifying those Responsible for Malicious Behaviour Online.* NATO Strategic Communication Centre of Excellence; 2022.
Reference Source
35. Margaros K, Paparrigopoulos T, Filippopoulos O, *et al.*: Detecting FIMI: A methodological framework for testing OSINT tools in TTPs detection. [Data set]. *Zenodo.* 2026.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ? ✓ ?

Version 2

Reviewer Report 26 August 2026

https://doi.org/10.21956/openreseurope.26801.r78801

© **2026 Dragomir M.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✓ **Marius Dragomir**
University of Santiago de Compostela, Santiago de Compostela, Spain

No more comments to make.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Disinformation, Media literacy, Media Studies, Broadcast regulation, PSM, Audience studies, Media sustainability,

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 20 August 2026

https://doi.org/10.21956/openreseurope.26801.r78802

© **2026 Spitale G.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Giovanni Spitale**
IBME, Universitat Zurich, Zurich, Switzerland

This is a serious revision. The authors have engaged with all eight major comments from my first report, and in several places they have gone beyond the minimum: the systematic-review framing has been withdrawn, the evaluation corpus is now described accurately as mixed-source rather than synthetic, the performance figures have been demoted from rankings to time-bound descriptive observations, and the ethics-by-design claim has been downgraded to a guiding orientation rather than defended as a measured outcome. That last move is uncommon in

revision and it improves the paper: where weakening a claim was the honest option, the authors weakened it. Appreciated. The new material in par 4.1 on content-dependent versus content-agnostic detection, the bias paragraph and operational safeguards in par 5.2, the decision protocol in par 5.2, the legal-versus-epistemic-ethics distinction in par 5.3, and the regulatory-access paragraph in par 6 address the substance of what I raised rather than gesturing at it.
My principal objection to version 1 was that the central empirical claim was not interpretable as stated and that the normative framework addressed lawfulness only. Both have been fixed properly rather than cosmetically, and the claims now match the evidence that supports them. I therefore move from Not Approved to Approved with Reservations.
The reservations that remain are of two kinds. Three concern the material deposited for this revision, which I checked against the article, and which does not currently do the work the article claims for it. The remainder concern residual sentences where the revised framing has not been propagated to text surviving from version 1. None require new data collection or re-analysis.

Disposition of the Version 1 comments
MaC1(a), circularity of the evaluation corpus for LLM testing: resolved. The first caveat in par 5.1 states the training-data recognition problem directly, par 3.2 repeats it, and the fact-checking paragraph now closes with the inference I said was missing, namely that the figures "should not be read as evidence that GPT-4o is the most reliable operational fact-checking instrument for novel, emerging, or out-of-distribution FIMI content". The study was not re-run on out-of-distribution material, but the paper no longer claims what such a re-run would be needed to support.
MaC1(b), inter-rater reliability: addressed by disclosure rather than by measurement. No coefficient was computed, and par 3.2, par 3.4, the fourth caveat in par 5.1 and the deposited README all say so. For a historical dataset this is acceptable, provided the consequence is stated, and par 5.1 now states it: the percentages are agreement with project-defined labels of unvalidated quality, not externally validated accuracy.
MaC1(c), the CT-A/CT-B divergence: resolved. The third caveat in par 5.1 names the 58 versus 95 percent gap explicitly and attributes it to the design rather than to the tool.
MaC1(d), the comparison as a validation strategy: largely resolved, with one gap. The clarification added at the end of par 3.3 is exactly right. It is not, however, applied at the point of use. The "Integrated and cross-domain evaluations" subsection still presents the most extreme cross-team contrasts in a comparative register, notably The Factual at "100 percent success in article analysis" for CT-A in the Disinformation Campaign Analysis and "0 percent success in article verification" for CT-A in the Cross-Platform analysis, set against CT-B's "more stable outcomes", with no reminder that these are not comparable quantities. One sentence at the head of that subsection referring back to par 3.3 would close this.
MaC2, systematic bias of LLMs on evaluative tasks: resolved. The paragraph added to par 5.2 is well judged. It names source-framing effects and broader ideological bias, correctly locates the mechanism in training-data and alignment regularities rather than in prompting, observes that GPT-4o is the least transparent instrument in the proposed suite, and makes its inclusion conditional on presenting content without source attribution and cross-validating against source-agnostic tools.
MaC3, legal compliance versus epistemic ethics: resolved. The new paragraph in par 5.3 draws the distinction and, importantly, connects it to the protocol in par 5.2 rather than leaving it as a standalone observation.
MaC4, the content-level detection paradigm: resolved. The closing paragraph of par 4.1 theorises the limitation as structural rather than contingent and distinguishes content-dependent

from content-agnostic approaches. See MiC3 for the one place where the conditioning is promised but not applied.
MaC5, normative architecture of the suite: resolved. The three-point protocol in par 5.2 is minimal but real, and it does the required work: verdict outputs as leads rather than conclusions, independent corroboration before investigative action, human override on divergence.
MaC6, the misinformation/disinformation conflation: partly resolved. The terminological paragraph in par 1 is clear and correct. It also makes a promise: "Where relevant, the analysis therefore distinguishes which layer of the tool suite addresses which type of epistemic harm". Sections 4 and 5 do not deliver this; the distinction is drawn once and then not used. Either deliver it, for which a clause per layer in par 5.2 would suffice, or soften the promise in par 1.
MaC7, the ethics-by-design claim: resolved. The second paragraph of par 3.5 is the honest version, and identifying the operationalisation of bias, transparency and explainability metrics as a priority for future iterations is the right disposition.
MaC8, the regulatory gap in future directions: resolved. The penultimate paragraph of par 6 names the question and identifies public-interest or institutionally guaranteed access to platform data as the most pressing open problem for the field.

Major Comments
MaC9: The Data availability section links to the superseded version of the Zenodo record (Data availability)
"Zenodo. Detecting FIMI: A methodological framework for testing OSINT tools in TTPs detection. https://doi.org/10.5281/zenodo.19045753"
That DOI is the version-specific identifier for version 1 of the deposit, published 16 March 2026, which contains two files: "Resonant documentation file.xlsx" (14.6 kB) and "Resonant Metrics.pdf" (601.6 kB). It is precisely the deposit that the Amendments note describes as insufficient: "the previous deposit contained only the metrics document and the documentation template". A reader following the citation arrives at a page headed "There is a newer version of the record available" and finds none of the corpus the article describes in Data availability.
The corpus was in fact deposited, as version 2.0, published 27 July 2026, 60.5 MB, comprising README.md, 1-4.zip, 5-8.zip, 9-12.zip and the two original documentation files. I verified its contents against the inventory given in the article and they correspond. The problem is one of citation only, and is trivially fixable, but until it is fixed the central transparency claim of this revision cannot be checked by any reader who follows the reference. Given that the deposit is what converts several of the version 1 objections from unresolved to resolved, this is not a cosmetic error.
RECOMMENDED FIX: cite the concept DOI 10.5281/zenodo.19045752, which always resolves to the latest version, or the version DOI 10.5281/zenodo.21627200. The same correction is required in Section 6 of the deposited README, which also cites the version 1 DOI. Two related items on the Zenodo side: the version 2.0 record's "Is supplement to" relation still points to 10.12688/openreseurope.21702.1 rather than to the current article version, and the record Description still lists only the template and the metrics document, with the corpus described in a separate Methods field, so a reader browsing the record sees the version 1 description attached to the version 2 files.
MaC10: The deposited README carries unexecuted pre-publication instructions, and provenance for the largest components is unconfirmed (Data availability, par 3.5, par 5.3)
The README is otherwise a good document, considerably better than what most method papers deposit, which is what makes this visible. Its file inventory carries repeated editorial notes addressed to the authors rather than to users: "Confirm original source and licence." against

fake.txt and real.txt; "Confirm source; treat as synthetic/illustrative unless otherwise stated." against bot_detection_data.csv; "Confirm source and redistribution licence." against asn-ipv4.csv and country_asn.csv. Section 5 then states: "Before publishing, the authors should confirm and record the exact source and licence for: the fake/real text corpora, the bot-detection dataset, the IP→ASN reference tables, the email corpus, and the website list."
That instruction was not carried out before the record was published. Three consequences follow, in ascending order of seriousness.
Epistemically, the unconfirmed provenance prevents a reader from evaluating the contamination caveat the authors added in response to MaC1(a). That caveat holds that GPT-4o "may achieve high apparent accuracy by recognising content encountered during training". Whether this risk is near-certain or negligible depends entirely on the origin of the 12,228 fake-labelled and 9,762 real-labelled text items. If they derive from one of the widely mirrored public fake/real news collections, contamination is close to guaranteed; if the material is project-generated, it is not. The README does not say, so the caveat cannot be assessed by the reader for whom it was written. Naming the source resolves this at no cost and would strengthen the caveat rather than weaken it.
Legally, the record is released under CC-BY 4.0 while redistributing components whose licences and database rights are, on the authors' own account, not established. The record's copyright note that third-party data is not re-licensed, and the README's assignment of compliance responsibility to re-users, mitigate this but do not resolve it: responsibility for observing a licence cannot be transferred for an act of redistribution that may not have been permissible in the first place.
Ethically and in data-protection terms, this is the point that bears most directly on the article's own commitments. The deposit redistributes email/*.eml, described in the README as 100 files with "real message headers", from a source recorded as a "third-party/public email corpus" with the licence unconfirmed, together with PhoneBookDataset.csv containing 441 rows of real contact data. An article whose par 3.5 invokes the project DPIA, purpose limitation and data minimisation, and whose par 5.3 argues that "public availability does not remove the need for appropriate handling", should not publish personal data whose lawful basis for onward publication is marked as pending in its own documentation. The README states the correct principle and the deposit does not yet meet it.
Recommended fix, in order of preference: confirm and record the sources and licences as the README itself instructs; where a source cannot be established, withdraw that component from the deposit and note the withdrawal in the README; and consider whether the email and phone-contact files need to be redistributed at all, since none of the article's claims depend on a reader possessing them. On any route, the "Confirm..." annotations should not remain in the published README.

MaC11: The instrument identified in the deposited baseline is not the one the article names (par 5.1, par 5.2)

The article refers throughout to GPT-4o. The deposited baseline document, Resonant Metrics.pdf, p. 3, lists the tool tested in the fact-checking category as https://chatgpt.com, and lists ChatGPT OCR separately in the multimedia category. If the testing was conducted through the consumer chat interface rather than through the API, two consequences follow that bear on interpretation. First, the evaluand is not the GPT-4o model but a product surface whose behaviour depends on the served snapshot, the system prompt and whether browsing and tool use were active, none of which the tester controls or observes. Second, if browsing was available, part of the reported accuracy may reflect retrieval of published fact-checks at query time rather than model knowledge or independent verification. That is a mechanism distinct from the training-data recognition

addressed in the first caveat of par 5.1, and it changes what the figures mean: a retrieval-assisted interface evaluated on documented cases behaves more like a search tool than like a classifier, and the resulting numbers say little about performance on content for which no published fact-check exists.

The article already concedes that "the exact historical GPT-4o snapshot, prompt, temperature and tool/browsing configuration were not retained". The present request is narrower and should be answerable from the testing partners' recollection: state which interface was used, and whether browsing was enabled, disabled, or not known. If the web interface was used, the instrument should be named accordingly throughout, for instance as ChatGPT (GPT-4o era, web interface), rather than as the model. This also affects par 5.2, where the recommendation to retain the tool in the operational suite is a recommendation about a specific access mode with specific safeguards attached.

Minor Comments

MiC1: The System Usability Scale statement contradicts the deposited documentation (par 3.4)

Par 3.4 states that "The retained records do not document a consistently administered System Usability Scale". Resonant Metrics.pdf, p. 8, states that "The System Usability Scale (SUS) was used to measure ease of use/user experience", lists three adapted items, and specifies that "Each team compiled only one response per tool based on its use in the various scenarios." Both statements can be true of the same facts, since a three-item adaptation is not the ten-item instrument, but as written the article and its own deposit contradict each other on their face. One clarifying sentence would resolve it, and should carry the detail that matters more than the label: usability was recorded as a single consolidated response per team per tool. That is a defensible design for an operational evaluation, but it bounds what "practitioner validation" can be taken to mean, and readers should be told in the article rather than left to discover it in the appendix.

MiC2: Percentages are reported without denominators anywhere in the article (par 5.1)

The article reports on the order of forty distinct percentages, most to two significant figures, without a single case count. Par 3.4 and par 5.1 explain that the complete denominator record is unavailable, and I accept that the run-level record is lost. But the presentation still asserts a precision the archive cannot support: "100 percent accuracy" for reverse image search, "16 percent" for Know News, "42 percent" for The Factual in the multi-source analysis, where the denominator could be five cases or five hundred. The deposit constrains this in places: the only reduced labelled text subsets it retains are 32 and 42 items, and the social-media examples are three files of five rows each. If the fact-checking figures rest on samples of that order, that is material to a reader assessing the CT-A/CT-B divergence, and approximate values are presumably recoverable from the partners even without the logs. Report approximate n where it can be established, even as a range; where it genuinely cannot, report coarse bands instead of two-significant-figure percentages. This concedes nothing the article can legitimately claim, and it removes an implied precision that the rest of the revision has been careful to disclaim.

MiC3: Residual sentences that contradict the revised framing

These are copy-level, but they matter, because a reader who meets them is entitled to infer that the qualifications elsewhere are pro forma.

Par 3.4: "Accuracy was examined by comparing the tool's outputs with the expected results embedded in the synthetic datasets." This contradicts par 3.2, which states that "Not all retained material was generated in-house or is synthetic", and it occurs in the description of the accuracy measurement itself.

Par 6, final sentence: "By offering a reproducible, ethically grounded, and operationally informed framework". This contradicts the first paragraph of the same section, "It therefore advances

methodological transparency without claiming a fully reproducible benchmark", and the Abstract. It is also the sentence most likely to be quoted.
Par 5.2, first sentence: "GPT-4o and The Factual were retained as the most accurate and scalable fact-checking instruments". The qualification arrives three paragraphs later, with the tool inventory in between, so a reader can take the recommendation and miss the condition. Add a forward pointer in the sentence itself.
Par 5.1, Disinformation Campaign Analysis: "the findings confirm that while content-verification tools are mature and reliable". Par 4.1 now closes by stating that this maturity claim "should be read as conditional on the tested content resembling previously documented cases", but the conditioning is only forward referenced from par 4.1 and never applied at the place where the claim is made.
Par 3, opening: "designed to provide a structured and reproducible process". The following sentences qualify it, but the contested word is doing unqualified work in the topic sentence.
Data availability, third paragraph: "The available data package has been deposited in Zenodo (link above and at [32])". Reference 32 is Germani and Spitale on source framing; the deposit is reference 35.
Par 6 places the incremental research agenda, that is machine-learning tools, pipeline orchestration and interoperability, immediately before the paragraph identifying the regulatory constraint as the more fundamental problem. Consider reordering so that the structural point is not read as an afterthought to the incremental one.
MiC4: Author contributions and acknowledgements
The Amendments note states that author contributions were added. I could not locate an author-contributions statement in version 2, and the Acknowledgements still read "The author would like to thank" and "the author wishes to thank" for a seven-author article. Worth confirming that the section survived production.

Recommendation: Approved with Reservations
The methodological contribution is real, and this revision brings the article's claims into line with the evidence that supports them. The two structural problems that made version 1 unacceptable have been addressed at the level at which they were raised, not deflected into limitations paragraphs. The "automation mentality" observation, which I described in my first report as the most accurate diagnosis in the article, now functions as part of its normative spine through the protocol in par 5.2 and the closing design principle in par 6, which is what I hoped would happen.
MaC9 and MaC10 should be treated as required corrections rather than as suggestions. MaC9 because the deposit on which the credibility of this revision rests is not the one the article links to, and no reader can currently verify the corpus claims by following the reference given. MaC10 because publishing personal data under an open licence while the lawful basis for redistribution is marked as pending in the authors' own documentation is inconsistent with the standard the article itself defends in par 3.5 and par 5.3, and because the unconfirmed provenance of the text corpora blocks assessment of the contamination caveat that resolves MaC1(a). Both are fixable without touching the analysis. MaC11 requires one clarifying sentence and possibly a renaming of the instrument. The minor comments are improvements rather than conditions.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: ethics, AI, disinformation, infodemic management

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 06 July 2026

https://doi.org/10.21956/openreseurope.23477.r74676

© **2026 Bozkurt V.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Veysel Bozkurt
Istanbul University, Istanbul, Turkey

Paper title: “Detecting FIMI: A methodological framework for testing OSINT tools in TTPs detection”
Reviewer recommendation: Approved with Reservations

The paper, developed under the Horizon Europe RESONANT project, presents a methodological framework for testing open-source intelligence (OSINT) tools aimed at detecting the tactics, techniques, and procedures (TTPs) used in FIMI (Foreign Information Manipulation and Interference) operations. The topic sits in an area that is highly current and carries real policy relevance. The lack of standardization in evaluating OSINT tools is a problem genuinely raised in the literature, and the authors' effort to approach it through a structured testing protocol deserves credit. Involving law enforcement partners from Greece and Ukraine in the testing process gives the study an operational grounding.
The research problem is clearly defined. The observation that OSINT tool evaluation remains fragmented and unstandardized is accurate and matches practitioner experience in the field. The three-layered detection logic built on the DISARM, AMITT, and MITRE ATT&CK taxonomies (content-level verification, platform-level monitoring, infrastructure-level attribution) offers an analytically functional framework. Using synthetic data instead of live operational data is the right choice on ethical and legal grounds. The references to GDPR and DPIA procedures are handled with care. Having law enforcement experts assess usability and interpretability alongside technical accuracy, rather than accuracy alone, is a genuine contribution that sets this study apart from purely academic reviews. The honest reporting of the structural fragility caused by tool discontinuation, API restrictions, and platform policy changes is a valuable warning for researchers in this field.
At the same time, the paper does not yet meet the core obligation expected of a method paper, namely describing the method at a level that others can apply and verify. The size, language, and construction procedure of the synthetic datasets are not reported. The performance indicators are defined conceptually but not operationalized at the level of measurement. Percentage values are presented without denominator information. The large discrepancies between the results of the

two testing teams are left unexplained. Until these gaps are addressed, the article's claim to reproducibility remains unsupported.

There is an internal inconsistency in the test design. Sections 3.2 and 3.3 state that each tool was exposed to identical inputs within its category, and that this is what makes the comparison meaningful. The opening paragraph of Section 5, however, states that the CT-A and CT-B teams tested the tools on different datasets and under their own operational conditions. These two statements contradict each other. The authors need to clarify the test design. Which tool was tested with which dataset, on how many observations, and by which team. Without this information, none of the findings in Section 5 can be properly assessed.

The synthetic datasets are not described. The paper places synthetic datasets at the center of its methodology but provides almost no technical information about them. How many claims, images, accounts, and email addresses were generated. How were ground truth labels assigned. What language were the datasets in. How was difficulty calibrated. None of these questions is answered in the text. The language question is especially important, since the authors' own findings show weak multilingual processing capacity. Whether the test data was in English, Greek, or Ukrainian directly affects how the results should be read. The 100 percent accuracy achieved in reverse image search also points to a ceiling effect, suggesting the dataset may lack discriminative power and be too easy. The inability to share the datasets for security reasons is understandable, but nothing prevents reporting descriptive statistics and the construction protocol.

The performance indicators are not operationalized. Accuracy, usability, interoperability, coverage, and sustainability are referred to as KPIs, yet none of them is defined at the level of measurement. Percentage values are given without denominators; a reader cannot tell whether a 90 percent accuracy figure comes from 10 observations or 500. There is no confidence interval, margin of error, or statistical test of any kind. No structured instrument (such as the System Usability Scale) was used for the usability assessment; the observations rest on free-form notes. Inter-rater reliability between the two teams was not calculated. A method paper exists precisely to document this kind of measurement protocol. The authors could prepare a table that specifies, for each KPI, an operational definition, a unit of measurement, and a data collection instrument.

The contradictory findings across teams are left unexplained. In the CT-A evaluation, GPT-4o correctly classified 58 percent of false claims; in the CT-B evaluation, this figure was 95 percent. The 37-point gap between the two is left unexplained. Because the comparability of the datasets and the implementation protocols across teams is not documented, this 37-point discrepancy cannot be interpreted.

The risk of training data contamination in the GPT-4o evaluation is overlooked. Declaring GPT-4o the most successful verification tool raises a serious validity concern. The synthetic claims were derived from known disinformation cases. There is a strong likelihood that these cases, along with the debunking content that refutes them, appear in GPT-4o's training data. What is being measured may therefore not be the tool's verification capacity but its capacity to recall training data. No conclusion about GPT-4o's superiority can be drawn without addressing this contamination risk. In addition, the test date, model version, prompt text used, and parameters such as temperature are not reported. Since large language models change behavior across version updates, this information is essential for reproducibility. It would also help to address the risk of hallucination and the limits of the model's ability to cite sources.

The tool selection protocol is not transparent, and the claim of being “systematic” is not supported. The abstract uses the phrase “systematic multi-source collection,” and the introduction claims to offer “the first systematic state-of-the-art review.” Yet the text contains none of the elements expected of a systematic review, such as search strings, databases searched, inclusion and exclusion criteria, or a flow diagram. How the tools were found and how they were narrowed

down cannot be traced. The selection protocol should either be reported transparently, or the label “systematic” should be removed from the text.
There is tension between the reproducibility claim and the data access policy. The article presents reproducibility as one of its main contributions. At the same time, the datasets, scenarios, and testing materials are kept closed for security reasons; the Zenodo record contains only non-sensitive supporting documents. The rationale for restricted access may be legitimate, but in that case the scope of the reproducibility claim should be narrowed and reframed around a more defensible notion, such as the transferability of the methodological template. As it stands, there is a tension between the claim and the data access policy that readers will notice. Adding at least the dataset schemas, size information, and measurement protocol to the Zenodo record would considerably ease this tension.
Conclusion and Recommendation
The article addresses a significant practical problem and occupies a distinctive position through its operational partnerships. However, it has not yet reached the core purpose of a method paper, which is to offer a transferable and verifiable protocol. The manuscript cannot be approved in its current form until the internal inconsistency in the test design is resolved, the datasets and measurement protocol are documented, and the contamination risk in the GPT-4o finding is addressed. That said, all of these issues are correctable, and the core idea is worth preserving.
My recommendation is Approved with Reservations. Once the points raised above are reasonably addressed, or the authors provide reasonable justification that satisfies the editor, I believe the article will reach an approvable standard. A further round of review should not be necessary. If the editor finds the revision satisfactory, the article can be accepted for indexing. I wish the authors success in the revision process.

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
No

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Sociology of Work/Social Impacts of AI

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 17 June 2026

https://doi.org/10.21956/openreseurope.23477.r75053

© **2026 Dragomir M.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Marius Dragomir
University of Santiago de Compostela, Santiago de Compostela, Spain

This article presents a methodological framework, developed within the Horizon Europe RESONANT project, for testing open-source intelligence (OSINT) tools used to detect tactics, techniques, and procedures (TTPs) associated with Foreign Information Manipulation and Interference (FIMI). The article reports per-tool performance observations, proposes an "optimized tool suite," and reflects on cross-cutting limitations including API volatility, weak interoperability, limited multilingual coverage, and the legal and ethical constraints of OSINT work. The topic is timely and important, the conceptual framing is appropriate, and the involvement of operational partners gives the study practical relevance beyond a purely academic tool review.

The rationale for the work is clearly and convincingly explained, which is why I have answered "Yes" to the first question. The authors establish that OSINT tool evaluation in the FIMI domain is fragmented and lacks standardization, and they motivate the shift from narrative-level analysis to TTP-oriented detection well. The grounding in DISARM, AMITT, and MITRE ATT&CK is appropriate, and the relevance to researchers, law enforcement, and EU policy audiences is articulated effectively. No revision is required on this point, in my view.

I have answered "Partly" to whether the method is technically sound because the overall logic is coherent, but the specification has gaps that I believe should be closed. The "Waterfall model" is invoked as the methodological inspiration but functions largely as a label; the authors should either demonstrate concretely how the stepwise structure shaped the testing design and traceability, or reduce this framing, since as written it adds citations without analytical payoff. More importantly, the operationalization of the TTP frameworks is underspecified: the authors should explain precisely how tools and test cases were mapped to specific DISARM, AMITT, and ATT&CK tactics and techniques. The manuscript also conflates three distinct claims that should be kept separate (whether a tool detects a TTP, produces evidence relevant to a TTP, or merely supports a broader investigative task), which represent very different evidentiary standards and should not be reported interchangeably. The evaluation indicators are also described conceptually rather than operationally, and there is a notable inconsistency between the manuscript and the project's own deposited materials: the article describes its indicators as accuracy, usability, interoperability, coverage of adversarial behaviors, and sustainability, whereas the deposited metrics file defines seven evaluation metrics on 1–5 ordinal scales (accuracy, coverage, speed/performance, ease of use, community feedback/satisfaction, customization/configurability,

and support/documentation) that do not cleanly map onto the five indicators named in the text, and do not clearly include interoperability or sustainability. The deposited materials further indicate that ease-of-use was captured through a System Usability Scale instrument with only one response per tool per team, so usability findings have no inter-rater basis unless an internal consensus procedure is described. The authors should reconcile the indicator sets, report results in the units actually used, state how each indicator was scored and by whom, and acknowledge the single-rater limitation.

I have answered "No" to whether sufficient details are provided to allow replication, and this is the most serious deficiency in my view. The paper claims a "systematic multi-source collection" of tools but provides no search strategy, source list, search dates, inclusion or exclusion criteria, number of tools screened versus retained, or rationale for excluding particular tools; as a result, the descriptor "systematic" is insufficiently documented (the authors may have followed a systematic process internally, but the article does not record it). They should add a transparent selection protocol (PRISMA-style or an equivalent structured account) and a final tool inventory with version numbers and testing dates. The synthetic datasets, which carry much of the study's validity, are described only in general terms: the authors should specify how each case was generated, how cases were mapped to specific TTPs, how realism and difficulty were calibrated, whether any real-world FIMI incidents served as anchors, whether domain experts validated the cases and annotations, and how many cases were used per category. The testing procedure also needs detail, and here the manuscript appears internally inconsistent: the methodology states that tools within a category were applied to the same dataset, implying category-level comparability, whereas the results section states that CT-A and CT-B each tested tools on distinct datasets under their respective operational conditions. The authors should clarify whether the two cyberteams tested identical or distinct datasets, since the reported percentages cannot be directly compared if the inputs differed. The procedure should also state how evaluator observations were captured and whether a shared rubric, consensus procedure, or measure of inter-rater agreement was used. In the publicly available Zenodo materials I examined, the open files document the intended evaluation framework and tool inventory but not the completed testing records; the logging template is empty. The deposit therefore documents the intended method without evidencing its execution and cannot substitute for the procedural detail requested here. These points should be addressed for the article to be scientifically sound as a method article.

I have answered "Partly" to the availability of source data, having examined the open Zenodo deposit (DOI 10.5281/zenodo.19045753. In the publicly available materials I examined, the deposit contains two files: a metrics document defining the evaluation indicators and their 1–5 scoring scales, together with the list of tools and the metric categories intended for each tool; and a documentation template in spreadsheet form. This makes the evaluation instrument and the tool inventory transparent, which is welcome and should be credited. However, the deposit does not contain the data underlying the reported results. The documentation spreadsheet appears to be an empty template, its logging sheet (columns for scenario number, tool, dataset, testing date and time, observed output, and issues faced) and its usability questionnaire are unfilled, and the synthetic datasets that generated the reported figures are, by the authors' own statement, restricted for security reasons. Consequently, none of the quantitative results presented in the article, namely the per-tool accuracy percentages, can be traced to deposited source data. A further misalignment compounds this: the deposited metrics file defines seven evaluation metrics on 1–5 ordinal scales, but these do not fully align with the five indicators described in the manuscript, and the article's results are reported mainly as accuracy percentages rather than in

the metric units defined in the deposited framework. To address the reproducibility claim, the authors should deposit the completed logs and scored results (suitably anonymized), reconcile the reported metrics with the framework actually used, and either release a sanitized subset of the synthetic datasets with their per-case TTP mappings or temper the reproducibility language accordingly.

I have answered "Partly" to whether the conclusions are adequately supported. The qualitative conclusions (that no single tool or category is sufficient, that content-verification and infrastructure-verification tools are relatively mature, and that bot detection and social media analytics are structurally fragile) are reasonable and broadly consistent with the reported observations. The difficulty lies with the quantitative claims and the tool recommendations. The results report many precise percentages, but the denominator is rarely given, ground-truth establishment is unspecified, and false-positive and false-negative definitions are absent; the authors should state, per category, the number of test cases, how ground truth was set, and how errors were classified. Several figures cannot currently be interpreted—for instance, an email-verification tool identifying "21–22 percent of valid addresses" is described as the most consistent performer, which is uninterpretable without the denominator and a definition of success. This reporting issue needs to be addressed, and it is compounded by the fact that none of these figures appears in the deposited documentation, and none is expressed in the 1–5 metric framework the project defined. A specific case requiring attention is the evaluation of generative AI systems: the article reports GPT-4o as a high-performing fact-checking instrument (around 90–95 percent accuracy), but where such systems are evaluated, the authors should report the exact model version, access date, prompt protocol, settings, retrieval or browsing conditions, number of runs, and the procedure for handling output variability; without these, the results are not independently reproducible, particularly as model behavior and product configuration can change over time. Finally, the proposed "optimized tool suite" includes commercial, proprietary, and platform-dependent tools whose functionality changes rapidly, and the selection criteria are not always transparent; the authors already acknowledge that results are time-bound, and they should make this caveat explicit in the recommendations themselves rather than presenting the suite as generally valid beyond the testing window.

In summary, the article makes a genuine and policy-relevant contribution and rests on a strong rationale, but it is not yet sufficiently specified to support its central claim as a reproducible methodological framework. The essential revisions concern transparency and reporting rather than the underlying work and are probably achievable quite easily.

The points that should be addressed in my view to make the article scientifically sound are those underlying the "No" answer on replication and the performance-reporting gaps under the conclusions question; on source data, the open deposit provides the instrument and tool list but none of the completed results or datasets that underlie the reported figures, and the reported percentages are neither documented in the deposit nor expressed in the metric framework the project defined.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?
No

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Disinformation, Media literacy, Media Studies, Broadcast regulation, PSM, Audience studies, Media sustainability,

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 15 June 2026

https://doi.org/10.21956/openreseurope.23477.r74680

© 2026 Spitale G. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Giovanni Spitale
IBME, Universitat Zurich, Zurich, Switzerland

The paper presents a structured methodology for evaluating OSINT tools against FIMI-related TTPs, developed within the Horizon Europe RESONANT project. The study tests eight tool categories using synthetic datasets and law enforcement practitioners as evaluators, mapping outputs to DISARM, AMITT, and MITRE ATT&CK frameworks. The methodological contribution is genuine: reproducible, ethically compliant testing of a fragmented tool ecosystem is a real gap, and the paper fills it competently. However, several structural problems undermine the reliability of the central empirical claim, the scope of the ethical analysis, and the normative adequacy of the proposed tool suite. These are detailed below.

Major Comments

MaC1: GPT-4o performance figures are not interpretable as stated (par5.1, fact-checking

paragraph)
"GPT-4o consistently outperformed all other tools, correctly identifying around 90 percent of true and 58 percent of false claims in the CT-A tests, and achieving 94 percent accuracy on true and 95 percent on false claims in the CT-B evaluation"
This is the most consequential empirical finding in the paper, and it is reported without the information needed to assess it. Three problems:

(a) The synthetic dataset is circular with respect to LLM evaluation. The authors describe synthetic content as "fabricated narratives derived from known disinformation cases" (par 3.2). GPT-4o was trained on exactly this type of known, publicly documented disinformation. The model's high accuracy on these inputs does not necessarily demonstrate fact-checking capability; rather, it might demonstrate recall of training data. The test would need to be conducted on novel, out-of-distribution content (information voids, emerging narratives, non-anglophone campaigns) to support the claim that GPT-4o "represents the most reliable fact-checking instrument." This distinction is acknowledged obliquely in par 6: "reliance on synthetic datasets [...] cannot fully capture the dynamic unpredictability of evolving information manipulation tactics", but the implication for the par 5.1 performance figures is not drawn. The conclusion in par 5.2 that GPT-4o should be "retained as the most accurate and scalable fact-checking instrument" follows only if the synthetic dataset is a valid proxy for operational conditions. It is not, I believe, and the paper should say so explicitly here rather than in a limitations paragraph that is structurally separated from the results. I recommend either to re-run the study on meaningful data or at least add a paragraph in par 5.1 explicitly addressing the circularity risk for LLM-based fact-checkers tested on synthetic data derived from known cases.

(b) No inter-rater reliability is reported for ground truth labelling. The synthetic dataset was produced internally by the RESONANT project (par 3.2). No information is given about how ground truth labels were assigned or validated. The accuracy figures in par 5.1 are ratios against these labels. Without inter-rater reliability measures, the accuracy figures are uninterpretable. I recommend reporting ground truth labelling procedure and inter-rater reliability.

(c) The large performance gap between CT-A and CT-B for the same tool (GPT-4o: 58% false-claim detection in CT-A vs. 95% in CT-B) is unexplained. A 37-percentage-point difference on false claims for an identical tool suggests that either the two synthetic datasets are not equivalent, or the evaluation criteria differed. The paper notes that teams "tested tools on distinct datasets under their respective operational conditions" (par 5 preamble) but does not analyze or explain what drives this divergence or what it implies for the methodological comparability the study claims. I recommend explaining the CT-A/CT-B performance divergence for GPT-4o or acknowledge it as a structural limitation that bounds the reliability of the comparative rankings.

(d) The CT-A/CT-B comparison is underpowered as a validation strategy (par 5 preamble, par 5.1) Using two teams on distinct datasets as a validation strategy conflates two sources of variance: team-level variation (skills, interpretation, operational procedures) and dataset-level variation (different synthetic content). The 100% vs. 43% accuracy difference for Search by Image in the Disinformation Campaign Analysis (par 5.1) and the 0% vs. 100% difference for The Factual in the Cross-Platform analysis are not interpretable at all under this design. The study cannot distinguish whether these differences reflect tool instability, dataset differences, or evaluator differences. This does not invalidate the methodology, but it means the performance figures in par 5.1 cannot be treated as reliable performance estimates, only as existence proofs that tools can produce these

outputs under at least one configuration.

MaC2: Known systematic bias of GPT-4o on information evaluation tasks is not acknowledged (par 5.1, par 5.2)
"GPT-4o and The Factual currently represent the most reliable fact-checking instruments" (par 5.1)
"GPT-4o and The Factual were retained as the most accurate and scalable fact-checking instruments" (par 5.2)
The recommendation to integrate GPT-4o as the primary fact-checking instrument in a FIMI detection suite for law enforcement is made without any reference to the documented systematic biases of large language models in exactly this evaluative task.
References demonstrated that the declared source of a text, independent of its content, systematically shifts LLM credibility scores. This bias is structural and reflects regularities in training data and cannot be corrected by prompting alone. Similar ideological and political biases have been widely documented in literature; https://link.springer.com/article/10.1007/s42001-025-00376-w; https://news.mit.edu/2024/study-some-language-reward-models-exhibit-political-bias-1210 ; https://cs191w.stanford.edu/projects/Huang,%20Flora_CS191.pdf ; https://aclanthology.org/2024.emnlp-main.508/ ; https://dl.acm.org/doi/abs/10.1145/3461702.3462624 ; https://www.degruyterbrill.com/de/document/doi/10.1515/jtc-2023-0019/html ; https://arxiv.org/abs/2405.16455). argues that the opacity of these biases precludes the calibrated trust that responsible use of any epistemic authority requires.

In the FIMI detection context, this has direct operational implications. Deploying a tool with structural known and documented source-attribution bias for attribution-related judgments is a threat to the validity of the detection paradigm the paper proposes. The authors mention in par 5.3 that "non-transparent scoring systems" reduce tool value, but do not apply this observation to GPT-4o, which is the most opaque tool in the entire suite.
IMHO Par 5.2 should include a paragraph on the source-framing bias documented in LLMs and its implications for GPT-4o's suitability as a neutral fact-checking instrument in geopolitically sensitive FIMI contexts. This does not require removing GPT-4o from the suite, but it requires qualifying the recommendation and specifying operational safeguards (e.g., instructions to evaluate content without source information presented to the model, cross-validation with source-agnostic tools, ...).

MaC3: The ethics section conflates legal compliance with epistemic ethics (par 3.5, par 5.3)
"All testing activities were conducted in line with the GDPR and under the project's DPIA. By using synthetic datasets and open-source tools, the methodology avoided high-risk data processing." (par 3.5)
"Analysts must ensure that their actions are necessary, proportionate and respectful of fundamental rights. This includes avoiding techniques that could be seen as intrusive, manipulative, or misleading." (par 5.3)
The ethical framework deployed in this paper is entirely juridical: GDPR compliance, proportionality, data minimization. This is appropriate for research ethics but insufficient for a paper that proposes deploying LLM-based verdict tools in law enforcement FIMI investigations. The paper does not engage with the distinct ethical problem of algorithmic judgment replacing human judgment in epistemic contexts. When GPT-4o flags a claim as likely false and this output feeds into a law enforcement investigation, the model functions as an epistemic authority. The opacity of that authority — its inability to provide reasons in principle open to scrutiny and

contestation — is an ethical problem that GDPR compliance does not address. The same paragraph that mentions "non-transparent scoring systems" as a usability issue (par 5.3) should recognize that non-transparency in verdict-producing tools is an epistemic ethics issue, not merely a UX issue.
Relatedly, the "automation mentality" paragraph in par 5.1 ("OSINT tools are simple, requiring only pressing a button and wait for the results [...] the tools can assist in the intelligence gathering process, but they cannot replace human creativity, analytical thinking, and reasoning") is the most intellectually honest passage in the paper, but it sits in isolation. It names the correct structural risk and then drops it. This risk should be integrated into the normative framing of the proposed tool suite in par 5.2 and the conclusions in par 6, not treated as a peripheral cautionary note.
I suggest that par 5.3 should distinguish between legal-compliance ethics (GDPR, DPIA) and epistemic ethics (the question of what kinds of algorithmic output are appropriate bases for investigative judgment). The "automation mentality" observation should be carried forward into the tool suite recommendations in par 5.2: which tools in the proposed suite produce verdicts, which produce structured evidence that analysts interpret, and what procedures govern when each type of output can be used operationally?

MaC4: The "content-level" detection paradigm's foundational limitation is not theorized (par 2.1, par 3.1, par 5.1)
"By focusing on the 'how' rather than the 'what' of disinformation, TTP frameworks enable analysts to identify recurring patterns, compare incidents, and design countermeasures." (par 2.1)
"Aligning tool functions with these frameworks ensured that the review addressed the full spectrum of TTPs relevant to FIMI detection" (par 3.1)
The paper correctly frames TTPs as structural rather than content-dependent at the taxonomic level (DISARM, AMITT). But the actual tools evaluated (fact-checkers, credibility databases, bot detectors) are almost entirely content-dependent in operation: they match specific claims against existing records, assess known domains, or detect known behavioral signatures. The paper notes that "novel narratives emerge" as a limitation of fact-checking tools (par 4.1) but does not recognize this as a fundamental paradigmatic limitation of the entire content-level detection layer. Research on viral misinformation detection has demonstrated that content-based systems fail systematically on information voids, precisely the events (crises, novel geopolitical developments) where FIMI operations are most dangerous and most active. The failure is not contingent on dataset quality or tool maintenance; it is structural. A methodology that presents GPT-4o and The Factual as the best fact-checking instruments must acknowledge that both tools depend on having encountered the relevant claims (in training data or indexed databases) before they can evaluate them. For genuinely novel FIMI content, the entire content-verification layer of the proposed suite degrades to near-zero.
The paper's conclusion that "content-verification tools are mature and reliable" (par 5.1, Disinformation Campaign Analysis paragraph) should be qualified to specify the conditions under which this holds.
Par 2.1 or par 4.1 should include a paragraph contrasting content-dependent and content-agnostic detection paradigms in case you feel you need theoretical backing on content agnostic paradigms. The structural limitation of content-based approaches for novel-content scenarios should be explicitly stated in par 5.1 and carried forward into the future research directions in par 6.

MaC5: The proposed tool suite has no normative architecture (par 5.2)
"No single tool or category was sufficient to address the full spectrum of adversarial tactics,

techniques, and procedures. However, by combining high-performing instruments across content, platform, and infrastructure levels, it is possible to build a detection ecosystem that compensates for individual weaknesses." (par 5.2)
The optimised suite is presented as a technical aggregation of best-performing tools across layers. What is missing is any normative specification of how the suite should be operated, in particular, what procedural constraints govern the use of verdict-producing outputs (GPT-4o's fact-check verdicts, credibility scores) as inputs to investigative or attributional processes.
The distinction between tools that produce structured evidence for analyst interpretation versus tools that produce verdicts intended for direct use is consequential in a law enforcement context. The paper treats this as a usability question ("unclear user interfaces, non-transparent scoring systems") but it is operationally and legally a question about the weight that different output types can carry. The absence of a proposed decision protocol (for example, specifying that GPT-4o fact-check outputs should require confirmation from a second independent source before informing any investigative action) means the "optimised suite" is a list of tools, not a framework for their use.
I believe par 5.2 should include at least a brief decision protocol specifying, at minimum: (a) which tool outputs are to be treated as evidence for further investigation versus conclusions; (b) which outputs require cross-validation before operational use; (c) conditions under which human review must override automated outputs.

MaC6: The "FIMI" label conflates analytically distinct phenomena (par 1, throughout)
The paper uses FIMI as a unified analytical category throughout. Within the detection context, however, the distinction between misinformation (false content produced without intent to deceive) and disinformation (false content produced with strategic intent) has direct operational implications: they require different attributional procedures and map differently onto legal responsibility frameworks. The paper acknowledges this implicitly (the TTP-based approach focuses on how operations are conducted, which is closer to disinformation (intentional, strategic)). But fact-checking tools (par 4.1, par 5.1) detect what is false, which is intent-agnostic. The conflation leads to inconsistencies: in par 3.2, the synthetic datasets include "fabricated narratives" (disinformation-type) alongside what appear to be simple false claims tested against fact-checking repositories (misinformation-type). Clarifying which part of the tool suite addresses which type of epistemic harm would sharpen both the methodology and the conclusions.

MaC7: The "ethics by design" claim is asserted but not operationalized (par 3.5)
"The approach was also guided by ethics-by-design principles, ensuring that AI- and NLP-based detection methods were assessed with attention to transparency, explainability and potential bias."
This sentence appears in par 3.5 but has no corresponding operationalization anywhere in the results or discussion. No bias assessment is reported for any AI-based tool. No transparency metric is defined or measured. No explainability criterion appears in the KPIs described in par 3.4, which cover accuracy, usability, interoperability, coverage, and sustainability. The "ethics by design" reference reads as rhetorical assurance rather than a methodological commitment. Either the bias/explainability assessment should be reported as a result, or the claim should be removed from par 3.5.

MaC8: Future research directions miss the most pressing structural gap (par 6)
"Future studies could expand the testing protocol to integrate machine-learning-driven tools, automated pipeline orchestration and cross-platform detection capabilities." (par 6)

The future research agenda proposed in par 6 extends the existing paradigm (more automation, more tools, more interoperability). It does not address the structural problem the paper's own results document: the two most critical detection categories (i.e. bot detection and social media analytics) are essentially non-functional due to platform API restrictions, and no technical solution to this is conceivable within the current third-party tool model. The real structural question, I.e. whether FIMI detection at scale requires public regulatory access to platform data, not better third-party tools, is not even named. The conclusion that these categories require "renewed development of sustainable, API-driven, and institutionally supported solutions" (par 5.1) is correct but stops short of the regulatory implication.

Recommendation: (very) Major revision
The paper addresses a genuine methodological gap: the OSINT tool ecosystem for FIMI detection is fragmented, and structured reproducible evaluation frameworks are scarce. The layered approach (content, platform, infrastructure), the practitioner-validated testing environment, and the explicit alignment with established TTP taxonomies are real contributions.
However, the paper in its current form cannot be accepted for two compounding reasons. First, the central empirical claim (that GPT-4o represents the most reliable fact-checking instrument and should anchor the proposed detection suite) rests on performance figures that are not interpretable as stated (MaC1: synthetic dataset circularity, absent inter-rater reliability, unexplained 37-point CT-A/CT-B divergence) and ignores a body of published evidence on systematic LLM bias in exactly the evaluative context the study proposes to deploy it (MaC2). Second, the normative framework is structurally incomplete: the ethics sections address GDPR compliance exclusively, while the paper proposes deploying verdict-producing algorithmic tools in law enforcement investigations without any decision protocol governing their use (MaC3, MaC5). The "automation mentality" paragraph in par 5.1 correctly names the core risk and then abandons it; this is the most accurate diagnosis in the paper and it should be its normative spine, not a footnote.
The remaining comments (MaC4 on content-dependence, MaC6 on the mis/disinformation conflation, MaC7 on the unoperationalized ethics-by-design claim, MaC8 on the regulatory gap in future directions) are substantive but addressable without new data collection. MaC1 and MaC2, however, require either re-analysis or a fundamental re-scoping of what the performance results are claimed to demonstrate.
The authors should be encouraged to revise. The methodological skeleton is sound, the empirical work is real, and the operational context (law enforcement, Horizon Europe) gives the paper genuine policy relevance. But it needs to be honest about what its data can and cannot support, and it needs to take the ethics of algorithmic judgment in investigative contexts as seriously as it takes GDPR compliance.

References

1. Germani F, Spitale G: Source framing triggers systematic bias in large language models. *Science Advances*. 2025; **11** (45). Publisher Full Text
2. Rozado D: The political preferences of LLMs. *PLOS ONE*. 2024; **19** (7). Publisher Full Text
3. Rozado D: The political preferences of LLMs. *PLOS ONE*. 2024; **19** (7). Publisher Full Text
4. Rettenberger L, Reischl M, Schutera M: Assessing political bias in large language models. *Journal of Computational Social Science*. 2025; **8** (2). Publisher Full Text
5. Liu Z: Cultural Bias in Large Language Models: A Comprehensive Analysis and Mitigation Strategies. *Journal of Transcultural Communication*. 2025; **3** (2): 224-244 Publisher Full Text
6. Boisseau É: Expertise, opacity, and trust in AI systems. *Synthese*. 2026; **207** (3). Publisher Full Text

7. Redaelli S, Spitale G, Germani F: The structural fingerprints of disinformation: a content-agnostic framework. *Ethics and Information Technology*. 2026; **28** (2). Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
No

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: ethics, AI, disinformation, infodemic management

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 05 June 2026

https://doi.org/10.21956/openreseurope.23477.r75059

© 2026 Jaidka K. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Kokil Jaidka
National University of Singapore, Singapore, Lower Kent Ridge Rd, Singapore

This article presents a methodological framework for evaluating OSINT tools for the detection of tactics, techniques, and procedures associated with Foreign Information Manipulation and Interference. The framework is developed within the RESONANT project and combines tool mapping, synthetic datasets, controlled testing by practitioner teams, and evaluation across dimensions such as accuracy, usability, interoperability, coverage, and sustainability.

My assessment focuses on the methodological, conceptual, and socio-technical dimensions of the study. The article has several strengths:

- The rationale for the study is clear and timely. The paper identifies an important gap in the FIMI and OSINT ecosystem: although many tools exist for fact-checking, multimedia verification, bot detection, social media analysis, and infrastructure-level lookup, their evaluation remains fragmented and insufficiently standardised. The move from narrative-level disinformation analysis to TTP-oriented detection is also well motivated. The use of DISARM, AMITT, and MITRE ATT&CK as conceptual anchors is appropriate, and the article is relevant to researchers, law enforcement agencies, and EU policy audiences.

- It usefully organises the OSINT tool landscape into content-level, platform-level, and infrastructure-level detection layers. It also gives appropriate attention to practical constraints that often shape real OSINT work, including API volatility, discontinued tools, subscription barriers, limited multilingual coverage, lack of interoperability, and the need for human judgment. The involvement of law enforcement partners from Greece and Ukraine gives the study operational relevance and helps move the discussion beyond a purely academic tool review.

However, the manuscript requires substantial revision before it can support its central claims as a method article. My main concerns are:

- The method is described at a relatively high level but is not specified in enough detail for replication or independent assessment. The paper states that tools were collected through a systematic multi-source process, but it does not provide a sufficiently transparent search strategy, inclusion and exclusion criteria, number of tools screened, final tool list with versions, testing dates, access conditions, or rationale for excluding particular tools. Without these details, the tool review cannot be assessed as systematic or reproducible.

- The details on the construction and validation of the synthetic datasets are unclear. The paper states that the datasets included fabricated narratives, manipulated and AI-generated images, simulated bot activity, and spoofed infrastructure elements. This is a reasonable strategy for avoiding privacy, ethical, and operational risks. However, the manuscript does not provide enough detail about how these cases were generated, how they were mapped to specific TTPs, how realistic or difficult they were, whether experts validated them, or whether any real-world FIMI cases were used as anchors. Synthetic datasets can support controlled evaluation, but their validity depends heavily on their realism, diversity, and annotation quality.

- The performance analysis needs much stronger reporting. The paper reports many percentages for individual tools, including GPT-4o, The Factual, Google Images, OCR tools, Geo-IP tools, phone verification tools, and email verification tools. However, the denominator is often unclear. The paper should state how many test cases were used for each category, how ground truth was established, how false positives and false negatives were defined, and whether the same inputs were used across both cyberteams. It should also explain how evaluator observations were recorded and whether any inter-rater agreement, consensus procedure, or structured rubric was used. At present, it is difficult to determine whether the results are robust evaluation findings or descriptive observations from practitioner testing.

- The claim of reproducibility is not fully supported. The article states that the full synthetic scenarios, detailed dataset structures, and related testing materials are restricted. This may be justified for security reasons, but it creates a tension with the article's repeated claims that the framework enables reproducible testing. A method article should provide enough procedural detail for another team to understand and apply the method with minimal ambiguity.

- The distinction between method, results, and recommendations should be clarified. The article moves from tool descriptions to tool performance and then to an "optimised tool suite," but the basis for recommending particular tools is not always sufficiently transparent. This is especially important because some of the recommended tools are commercial, proprietary, subscription-based, dependent on platform access, or subject to rapid changes in functionality. The authors should avoid presenting tool rankings as generally valid beyond the specific testing period unless the evaluation design supports that claim.

The manuscript would also benefit from clearer operationalisation of the TTP frameworks. DISARM, AMITT, and MITRE ATT&CK are appropriate references, but the paper should explain more precisely how they were used in testing. Were tools mapped to specific tactics and techniques? Were synthetic test cases designed to correspond to particular TTPs? Did evaluators assess whether a tool could detect a TTP, produce evidence relevant to a TTP, or simply support a broader investigative task? These are distinct claims and should not be conflated.

In conclusion, this is a useful and policy-relevant article, but it is not yet sufficiently specified as a methodological framework. The rationale is strong, and the topic is important. However, the current version does not provide enough methodological transparency, source-data availability, or performance-reporting detail to allow independent replication or full assessment of the findings.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
No

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: misinformation, computational linguistics, discourse analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.