Towards a Trustworthy AI Solution for Disinformation Detection

Eleni Tsalapati[1], Giorgos Giotis[1], Iosif Lazaridis[1], Thanasis Tsafonis[1], Christina Vaia[1], Ilias Kanellos[1], Cristina Guseth[2] and Teodora Musat[2]

[1]*Athens Technology Center, Rizariou, 10, Chalandri, Greece*
[2]*Freedom House Romania, Bld. Ferdinand I, nr. 125, sector 2, Bucharest*

Abstract

The rapid spread of disinformation and hate speech across digital platforms highlights the urgent need for effective support mechanisms for media professionals and journalists. Artificial Intelligence (AI) tools can significantly enhance the detection and analysis of misleading or harmful content, enabling faster and more scalable responses. However, for such tools to be meaningfully integrated into journalistic workflows, they must demonstrate a sufficient degree of trustworthiness, including robustness, reliability, security, and transparency. Since achieving absolute trustworthiness in AI systems is unrealistic, systematic risk assessment methodologies are essential to identify, evaluate, and mitigate potential weaknesses. In this paper, we apply the FAITH AI_TAF_V1 methodology to perform an in-isolation risk assessment of the AI Intelligent Coaching application, a service designed for disinformation and hate speech detection. The study investigates the applicability and limitations of the methodology, while highlighting key trustworthiness challenges for AI-assisted journalism systems.

Keywords

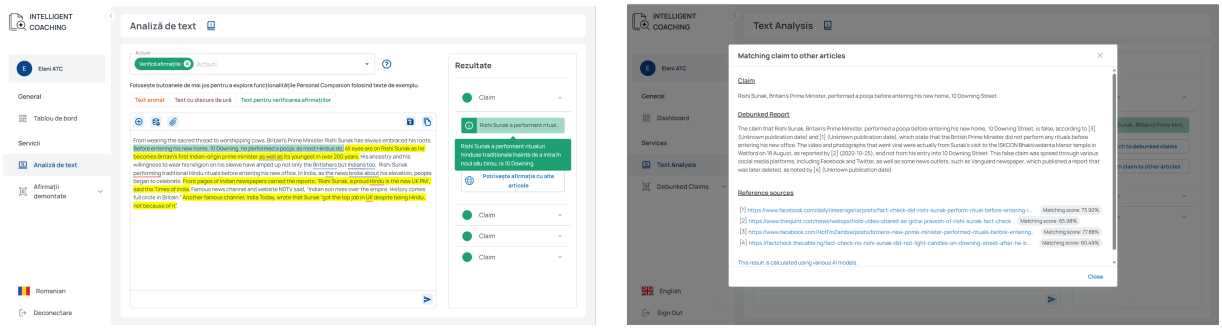
disinformation detection, large language models, trustworthy AI

1. Introduction

The European Union supports media freedom and pluralism as fundamental principles of modern democracy, fostering free and open debate. As noted by the High Representative for Foreign Affairs and Security Policy of EU in March 2026, "information manipulation is one of the most cost-effective tools of modern hybrid warfare". At the same time, as recently pointed out by the Director of the European Commission (EC) Department for Migration and Home Affairs (DG HOME), today, disinformation and hate speech are used to intensify crisis situations, aggravate divisions, cause political and even encourage physical harm. Beyond its societal importance, the detection of disinformation and illegal hate speech is increasingly a matter of legal obligation: under the Digital Services Act, providers of online platforms must operate notice-and-action mechanisms for illegal content, while very large online platforms and search engines are additionally required to assess and mitigate systemic risks such as disinformation campaigns and the dissemination of illegal hate speech (Articles 34–35), with the integrated Code of conduct on countering illegal hate speech and Code of Practice on Disinformation serving as benchmarks for compliance. So far, numerous initiatives, including past or ongoing EU-funded research projects, such as AI-CODE[1], vera.ai[2], TITAN[3], or AI4TRUST[4], have resulted in a wide range of AI-based solutions helping to combat disinformation and hate speech, especially via detection, and, therefore, in mitigating negative impacts on society. These research outcomes are currently being leveraged as

TRUST-AI: The Second European Workshop on Trustworthy AI. Organized as part of the International Joint Conference on Artificial Intelligence - IJCAI/ECAI 2026. August 2026, Bremen, Germany.
Corresponding author.
E.Tsalapati@atc.gr (E. Tsalapati); G.Giotis@atc.gr (G. Giotis); I.Lazaridis@atc.gr (I. Lazaridis); T.Tsafonis@atc.gr (T. Tsafonis); C.Vaia@atc.gr (C. Vaia); I.Kanellos@atc.gr (I. Kanellos); office@freedomhouse.ro (C. Guseth); office@freedomhouse.ro (T. Musat)
0000-0001-9464-404X (E. Tsalapati); 0000-0001-7125-9536 (G. Giotis); 0000-0003-2146-3795 (I. Kanellos)

[1]https://aicode-project.eu//
[2]https://www.veraai.eu/
[3]https://www.titanthinking.eu/
[4]https://pilot.platform.ai4trust.eu/



(a) (b)

Figure 1: (a) Claim Detection. The Intelligent Coaching application detects multiple claims. Upon selection, the decontextualized claim appears in a pop-up window (b) The "Match to Other Articles" functionality returns the decontextualized claim and a brief report generated from the detected resources, along with their URL and a matching score.

a backbone for novel AI-based services, which will be tailored for media professionals to help them reinforce trust in the content they produce.

Still, the real-world adoption of disinformation detection services depends on whether media professionals can trust them. Journalists must be confident that the tools they use are accurate, fair, robust, reliable, and secure. For example, claims identified as "debunked" should have indeed been fact-checked, while the system should avoid omitting relevant fact-checked claims within an article. It is equally important to ensure that AI-based tools are unbiased, do not discriminate against specific sources, and adequately protect user data. Trust can be further strengthened by providing explanations and justifications for the decisions generated by the AI system. However, due to the inherent interdependencies among trustworthiness characteristics—for instance, enhanced security mechanisms may reduce transparency—it is impossible to achieve absolute trustworthiness in AI systems. Therefore, assessing the trustworthiness of AI systems through a risk-based perspective is essential. FAITH AI_TAF_V1 [1] methodology constitutes a first attempt in this direction, enabling the in-isolation assessment of AI systems, i.e., without considering cascading effects across interconnected assets.

In this paper, we demonstrate the application of the FAITH AI_TAF_V1 methodology to the Intelligent Coaching application, an AI assistant tailored for media professionals. Developed within the AI-CODE project, the Intelligent Coaching application is a large language model (LLM)-based system designed to support journalists and fact-checkers in verifying the veracity of news articles. To evaluate the associated risks, media professionals from the Freedom House Romania network participated in a series of stakeholder workshops. Through these sessions, they assessed the business impacts of potential compromises in key trustworthiness characteristics, thereby highlighting the principal challenges to establishing user trust in such systems.

2. Intelligent Coaching Application

The AI Intelligent Coaching application aims to assist journalists in fact-checking the articles they write or read by checking the content's factuality and whether they contain certain types of logical fallacies or ch. In this paper, we specifically focus on the following two functionalities:

Claim Verification: It helps end-users assess the content factuality of an input text. This functionality consists of 3 modules:

- The Claim Detection module (Figure 1(a)) detects the core claims in the text and decontextualizes them (i.e., it adds from the input article all required information for the claim to stand on its own). This module consists of a pipeline of transformations (prompt chains) on the input text using LLMs, to finally derive the text snippets containing the claims.

- The Match to Debunked Claims (or Match to DB in short) searches the locally stored database with fact-checks (Debunked Claims database) if there is a fact-checking report that assesses a claim semantically similar to the detected claim selected for analysis and generates a short report based on the detected fact check. This module is based on a vector DB (which is the transformation of the Debunked Claims database into an embedding space), an API, and an AI Agent developed using LangGraph[5] and LangChain[6].
- The Match to Other Articles (Figure 1(b)) searches the Web for sources that debunk the detected claim selected for analysis and generates a short report based on the detected sources. The implemented workflow expects a claim (i.e., a sentence) as its input, then performs a web search across various search engines and web resources based on this claim and search keywords generated by generative AI. Then, it categorizes the collected results into different classes leveraging LLMs as classifiers [2].

Hate Speech Detection: Hate speech, often defined by defamatory remarks directed at specific groups based on factors such as ethnicity, nationality, religion, race, gender, and sexual orientation [3], poses significant challenges in media discourse. It is often that automated and AI-based detection fail to capture context and nuances in human communication. The Hate Speech Detection module performs a classification task split into two sequential prompts for improved accuracy even in highly contextualized cases[7]. The first prompt, a binary classification few-shot example instructs the model to determine whether the input contained hate speech. The second prompt is issued to classify the hate speech into the appropriate category (Racism, Sexism, Religious or Sexual Orientation).

All LLMs used are provided by the third-party service Groq[8]. Although LLMs have demonstrated impressive capabilities to perform knowledge-intensive tasks (e.g., [4]), they also face numerous challenges, for instance, they are hindered by hallucinations (e.g., [5], [6]), or, as recently found, they may solidify or exacerbate bias [7]. Hence, for none of the aforementioned tasks have LLMs been employed for the AI Intelligent Coaching service off-the-shelf. Instead, agenting approaches have been applied to overcome these drawbacks. Still, though, the limited number of available datasets, the general limitations of LLMs, and the continuous advancements in obfuscation techniques developed by disinformation spreaders introduce several risks. These risks are discussed in detail in Section 4.

3. Methodology

In this research work, the methodology FAITH AI_TAF_V1 [1], developed within the FAITH innovation action[9], is followed. With this methodology, the various assets of an AI system are assessed in isolation, i.e., without taking into consideration any cascading effects. This allows the study of the intrinsic trustworthiness properties of each individual AI asset. In future work, the AI Intelligent Coaching application will be assessed as a whole, by examining the interactions and interdependencies among the various AI assets.

As illustrated in Figure 3, the methodology consists of six phases. In the *Cartography Phase*, the boundaries of the assessment are clearly defined to establish the scope of the trustworthiness and risk analysis. For this case, the boundaries include the AI models: Claim Detection, Match to Debunked Claims, Match to Other Articles and Hate Speech Detection. All models (except for Match to Debunked Claims) have been designed and configured primarily through the use of zero-shot learning techniques, complemented by iterative user testing and validation during pilot activities. Additionally, the system is assessed for the *deployment* lifecycle stage.

[5] https://www.langchain.com/langgraph
[6] https://www.langchain.com/about
[7] More information about the performance of the Hate Speech Detector can be found in the report: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e51d5e7dd4&appId=PPGMS
[8] https://groq.com/
[9] https://faith-ec-project.eu/

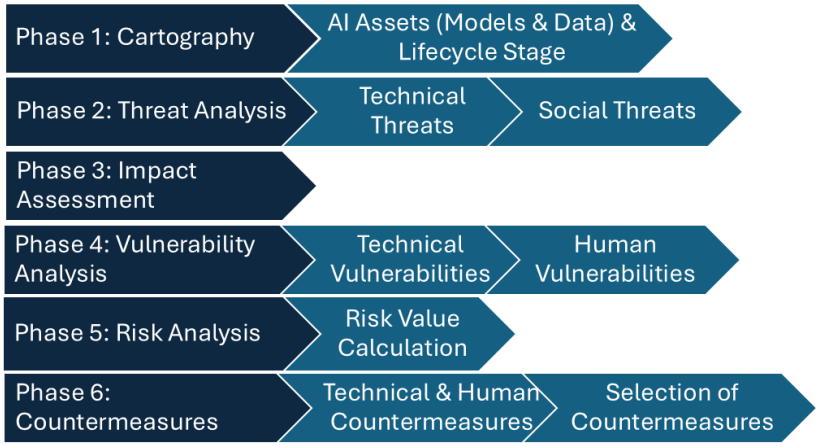


Figure 2: The six phases of the risk-assessment methodology based on FAITH AI_TAF_V1 [1].

For the *Threat Analysis*, a desk analysis is conducted to comprehensively examine and instantiate the diverse threats of all AI assets during each phase of the AI lifecycle that could undermine system trustworthiness. In this research paper, we focus on the threat taxonomy developed by Caroline et al. [8] to assess its applicability and limitations for LLM systems. Four criteria have been taken under consideration to assess a threat's likelihood: i) attacker's capability (e.g., the attack needs only commonly available skills or tooling vs it needs white-box gradients, sustained oracle probing, or privileged access to the local environment), ii) asset exposure (it is internet-facing or ingests untrusted external content vs is local, static, and isolated from untrusted input), iii) AI-Act criticality: The system is an assisting-only, low-criticality tool, so targeted attacks are judged improbable, and iv) Component nature (e.g., the component is generative/LLM-based, with a broad probabilistic surface and hence subject to prompt sensitivity, hallucination, oracle-style behaviour vs it is deterministic, with a narrow, predictable surface). Additionally, the ENISA/ISO-27005-style ordinal likelihood scale has been adopted.

The *Vulnerability Analysis* focuses on identifying weaknesses in the AI system. In this research paper, the weaknesses are identified through the corresponding controls defined by Caroline et al. for each of the identified threats in the threat assessment phase. The *Risk Analysis* calculates the various risks across the trustworthiness characteristics, based on the results from the impact, threat and vulnerability analysis, while in the *Countermeasures* phase the non-implemented controls are suggested to the user to enhance the trustworthiness of the AI system.

In this research paper, we focus on the first four phases and, following FAITH AI_TAF_V1, on the trustworthiness characteristics defined in NIST [9]: **Safety** mandates prevent any safety risks to humans, property, health, and the environment. **Security and resilience** require that the AI system can cope with unexpected adverse events or environment changes and, if unavoidable, it degrades safely and gracefully. **Explainability and interpretability**: Explainability refers to a "representation of the mechanisms underlying AI systems' operation". Interpretability refers to the "meaning of the output of AI systems in the context of their functional purposes designed". **Privacy-enhancement** can be intended as the overall "protection of human identity, autonomy, and dignity". **Fairness** requires ensuring an equitable development and use of an AI system, thereby avoiding harmful discrimination and biases. **Accountability and transparency** are coupled towards the assurance that a responsible approach to AI development is adopted and that all the relevant pieces of information about it are accessible to appropriate levels of the stakeholders, actors and users involved. Transparency reflects the "extent to which information about an AI system and its outputs is available to individuals interacting with such a system". Accountability is based on transparency. **Validation** is the "confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled" (Source: ISO 9000:2015). **Accuracy and robustness** contribute to the validity and trustworthiness of AI systems, and can be in tension with one another in AI systems.

4. Impact Assessment

The impact (or alternatively, consequence) assessment was conducted by FH based on insights collected through a series of 17 workshops involving a total of 75 media professionals. During these workshops, participants were asked to evaluate the potential business consequences arising from the compromise of the trustworthiness characteristics presented in Section 3 within the context of a disinformation detection service, such as the AI Intelligent Coaching application. For instance, an example question was *"What is the business consequence level caused by a privacy compromise of the AI system?"*. In the following, the trustworthiness characteristics associated with the highest assessed impact levels are discussed.

Privacy: Very high impact. In the context of journalism, a privacy compromise poses an especially serious risk because articles and written content that may contain sensitive investigative data, unpublished reports, or confidential documents are processed by the AI system. Unauthorized access to this kind of information could compromise journalistic privacy, harm one's reputation as a professional, and deter users from using the platform. The potential commercial repercussions of a privacy breach are thought to be extremely severe, given the significance of safeguarding editorial materials and journalistic sources. This evaluation is in line with the EU AI Act's strong emphasis on data protection and privacy safeguards, which emphasizes the necessity for AI systems to guarantee safe data handling, personal information protection, and suitable governance mechanisms throughout the system lifecycle. Additionally, it is consistent with the European Commission's more general guidelines for trustworthy AI, which emphasize the significance of privacy and responsibility as essential conditions for the ethical development and application of AI technologies.

Fairness: High impact. A crucial prerequisite for AI systems that analyze journalistic content is fairness. The interpretation of particular stories or subjects during the analytical process may be impacted if the system generates unfair or biased results. Such bias could damage the tool's trustworthiness in a media setting and create moral questions about how to understand delicate subjects. Even though journalists still have the ultimate authority over editorial decisions, chronic bias in AI results might undermine user acceptability and erode system trust. These factors make the potential business impact of inadequate fairness considerably high.

Accuracy and Robustness: High impact. The precision and resilience of the AI system's analytical results are critical to its efficacy. The system's ability to spot false information, identify misleading assertions, or flag dangerous content would be greatly diminished if it generated erroneous, inconsistent, or unstable results. Journalists may have to spend more time confirming results in these circumstances, which could reduce the tool's anticipated productivity gains. Additionally, persistent errors may erode user confidence and lower the possibility of long-term adoption. As a result, inadequate accuracy and robustness are thought to have a significant business impact.

5. Threat Analysis

The threats catalogued by Caroline et al. [8] are examined here for their applicability to, and limitations within, the AI Intelligent Coaching application. This assessment was conducted as a desk analysis by the development team and the risk assessor. Because operational fact-checking and disinformation-detection systems are an emerging class of application, documented incident data on which to base frequency estimates remain scarce; we therefore anchor each likelihood to the literature where relevant evidence exists—cited in Table 1—and, where it does not, derive it from the specific characteristics of the affected component and its deployment environment, according to the scale and criteria defined in Section 3.

Specifically, since this application acts only as an assisting and not a decision-supporting tool for journalists, the low-criticality classification of the application under the AI Act (an assisting, non-decision-supporting tool) sets an upper bound on all levels, For instance, the *Oracle* threat, in which the attacker explores a model by providing a series of carefully crafted inputs and observing outputs, has low likelihood in models such as Claim Detection or Match to DB because it requires sustained, high-effort probing of the system's input–

output behaviour to infer internal decision logic, which is rather unlikely due to its low criticality level. On the other hand, the Hate Speech Detection model is more susceptible to the oracle threats, because it deals with adversarial users actively trying to bypass rules leveraging automated tools (e.g., AutoDAN) [10].

Furthermore, asset exposure and component nature account for the spread between components. Thus, the deterministic, locally-hosted Match to DB module — with no reachable proprietary logic and no untrusted input — is assigned Never or near-negligible likelihoods, whereas the generative Match to Other Articles module, which ingests external web content through third-party search APIs, reaches more than once a year for the same threat classes. Externally hosted model internals are capped at once every 10 years, reflecting the rate-limiting and monitoring enforced by the API provider.

The *Data Disclosure* threat, which refers to a leak of data manipulated by ML algorithms due to factors like inadequate access control, third-party data sharing, or project team errors, has a low likelihood for the Claim Detection and Hate Speech Detection modules (occurring once every 5 years) because the third-party provider (Groq) implements reasonable and appropriate data encryption and security designed to protect customer data. Similarly, the Match to DB component has no likelihood of threat ("Never"), because the database is open to end users and the matching is a deterministic property of the stored embeddings, meaning there is no proprietary logic to "disclose" in this component. On the other hand, the Match to Other Articles module is much more susceptible to the data disclosure threat (occurring once a year), because it transmits proprietary search queries via Search APIs (Google/Tavily).

The *Model Disclosure* threat, which involves the leak of model internals or system prompts, has varying threat levels across components. The Hate Speech Detection module is the most susceptible (occurring more than once a year) because attackers can treat the binary classifier as an oracle to map safety boundaries and disclose hidden moderation logic. The Claim Detection and Match to Other Articles modules share a moderate likelihood (once every 2 years); the former is vulnerable to reverse-engineering via Direct Context Extraction, while the latter is susceptible to inference-based exfiltration, where iterative probing can force the model to unintentionally reveal hidden context or orchestration logic. Conversely, the Match to DB component has a very low likelihood because it utilizes standard RAG logic, making the model architecture a low-value target for disclosure compared to the data itself.

The *Model or Data disclosure* threat refers to the possibility of leakage of all or partial information about the model. The likelihood of this threat is from "Never" (Match to DB) to "Once every 10 years" (the rest of the models). This is mainly due to the use of externally hosted models. API providers implement strict rate limits and monitoring to prevent the high-volume probing required for model extraction. Similarly, the *Compromise of ML application components* threat, which refers to the compromise of a component or developing tool of the ML application, is also considered as of low likelihood. This threat is limited to the local environment's security. This is a threat to the AI Intelligent Coaching application only if targeted attacks are implemented to compromise the local SDL or the environment in which it resides, which considered highly unlikely.

Another threat of medium to low likelihood (once every 2 years) for the system under test is the "*Use of adversarial examples crafted in white or grey box conditions*", which is materialized when the attacker gains access to information (model, model parameters, etc.) that allows the direct built of adversarial examples. By using Llama models through Groq, the AI Intelligent Coaching application exists in a unique "Grey-Box" state. While Groq acts as a black-box API (preventing direct access to gradients), the Llama models used are open-weight, hence an attacker can download the exact same model weights from Meta/HuggingFace, craft a White-Box attack on their own machine using gradients, and then send that finished "adversarial payload" to the Groq-hosted system. Because the models are identical, the attack has a near-perfect Transferability rate. However, this type of attack requires high technical skill and must be targeted, which we consider highly improbable.

The threat of "*Unreported Cybersecurity Incidents*" varies significantly based on staff perception and operational pressure. The Hate Speech Detection module faces the highest risk (more than once a year) because staff frequently fail to report jailbreaks or filter bypasses if they believe the issue was eventually resolved or fear administrative blame [10]. The Claim Detection module has a moderate likelihood (once every 2 years), as high-pressure environments prioritize production speed, leading staff to dismiss minor extraction anoma-

lies as minor technical glitches. Conversely, the risk is lower for Match to Other Articles (once every 5 years) because staff may not recognize suspicious search behaviors or indirect prompt injections as active security threats. Finally, the Match to DB component has no likelihood ("Never") because its deterministic, internal nature means issues are classified as standard system failures and naturally reported straight to IT support.

The threats presented in Table 1 are the ones that are most relevant to the AI Intelligent Coaching as a whole. According to [8]; the threat "*Evasion*" refers to the type of attacks that target the ML algorithm's inputs to find small perturbations leading to large modification of its outputs (e.g., decision errors); the threat "*Failure or malfunction of ML application*" refers to failures, such as denial of service due to bad input or unavailability due to a handling error; the "*Denial of service due to inconsistent data or a sponge example*" refers to inappropriate input data formats or specifically designed input data to increase the computation time of the model; the "*Poisoning*" threats focus on the data or the model to modify the ML algorithm's behavior in a chosen direction (e.g. to sabotage its results, to insert a backdoor); finally, the "*Human Error*" refers to the various mistakes that can be made by the different stakeholders resulting in a failure or malfunction of ML application.

Threat	**Component**	**Likelihood & Description**
Evasion	Claim Detection	Once a year. Being a multistage pipeline system, evasions can happen in the summarization phase (summarization hijacking) or the ranking phase (ranking manipulation)[11].
	Match to DB	Never. This module performs a similarity check against a local database. Unlike a generative process or a search of the open web, the matching is a comparison between the user's claim and a static, pre-existing set of entries.
	Match to Other Articles	Once a year. As the system consumes external content via Search APIs (Google/Tavily), an attacker can perform Data Source Impersonation by publishing web pages containing Indirect Prompt Injections (IPIs).
	Hate Speech Detection	Once a year. Lexicon evasion using metaphors or dialects can bypass binary classification systems [12].
Failure or malfunction of ML application	Claim Detection	More than once a year. Large-scale summarization creates high KV cache demands; malformed inputs may cause memory overflow or node crashes [13].
	Match to DB	Once every 5 years. Failure is unlikely due to local static database; issues only arise from orchestration logic or DB timeout.
	Match to Other Articles	More than once a year. Highly dependent on external APIs; rate limits or latency spikes can propagate cascading failures in LangGraph (ASI08)[10].
	Hate Speech Detection	More than once a year. API saturation (e.g., Groq limits) may cause service unavailability.
Denial of service due to inconsistent data or sponge example	Claim Detection	More than once a year. Vulnerable to sponge examples and high-entropy inputs that saturate KV cache and cause timeouts [10].
	Match to DB	Once every 5 years. Static structured data reduces vulnerability to computational overload attacks[11].
	Match to Other Articles	More than once a year. Malicious web content (non-UTF-8, recursive HTML) may cause inference hang and GPU saturation.
	Hate Speech Detection	More than once a year. Sponge inputs can force excessive computation and slow system throughput [14].
Poisoning	Claim Detection	Once every two years. Semantic context deception may affect summarisation accuracy via poisoned external context [15].
	Match to DB	Once a year. Vulnerable to knowledge base poisoning via adversarial documents in vector DB [16].

[10] https://genai.owasp.org/2026/04/14/owasp-genai-exploit-round-up-report-q1-2026/
[11] https://genai.owasp.org/llmrisk2023-24/llm04-model-denial-of-service/

Threat	Component	Likelihood & Description
	Match to Other Articles	More than once a year. Web-scale poisoning via indirect prompt injection may alter ranking behaviour [17],[18].
	Hate Speech Detection	Once every 5 years. Resistant unless underlying model weights are compromised.
Human error	Claim Detection	Once a year. Prompt modifications without regression testing may degrade summarization or output formatting.
	Match to DB	More than once a year. Mis-indexing or incorrect labeling in local DB leads to false negatives.
	Match to Other Articles	More than once a year. Misconfiguration of search parameters may cause context overflow or hallucinated consensus.
	Hate Speech Detection	More than once a year. Misuse of module for unintended tasks (e.g., sarcasm detection) leads to misclassification drift.

Table 1: Threat likelihood assessment across system components based on expert desk analysis.

Besides the aforementioned threats, the literature indicates that LLM-based systems are susceptible to several additional types of threats (e.g., [19], [20]). Only OWASP[12], in 2025 introduced 17 new threats relevant to LLMs and agentic AI systems. Indicatively, some threats highly relevant to the AI Intelligent Coaching application include:

- Prompt Attacks: LLMs are highly sensitive to the structure of a prompt; even a minor modification by an attacker may lead to misleading responses.
- Human Attacks on Multi-Agent Systems: Attackers exploit inter-agent trust relationships and workflow dependencies to manipulate agent operations or escalate privileges.
- Overwhelming Human in the Loop: Attackers exploit human cognitive limitations or weaknesses in human oversight mechanisms.
- Misaligned and Deceptive Behaviors: Autonomous agents develop unintended or deceptive strategies that may lead to harmful outcomes.
- Intent Breaking and Goal Manipulation: Attackers manipulate an agent's goals, planning, or reasoning processes to redirect its behaviour.
- Cascading Hallucination Attacks: False AI-generated information propagates through the system, disrupting reasoning and decision-making processes.

6. Vulnerability Analysis

Technical vulnerabilities of the AI system, particularly those identified by Caroline et al. [8] that are associated with the threats exhibiting the highest likelihood, as well as with the most impacted trustworthiness characteristics identified in Section 4, are presented in Table 1. As discussed in Section 3, the vulnerability assessment phase of the methodology also considers human-related vulnerabilities; however, these are left for future investigation.

According to FAITH AI_TAF_V1, the vulnerability assessment process examines which of the proposed mitigation controls have been implemented. The relevant controls are presented in Table 2. They are partitioned into three groups: those applicable to developer-owned components, denoted with "A", those non-applicable as model-engineering controls, denoted with "NA", but recoverable as supplier-governance controls, and a residual core, closable only through supplier attestation. As shown in the table, a significant number of controls are considered non-applicable as model-engineering controls because the AI Intelligent Coaching application relies heavily on third-party services over which developers have limited control. For example, controls related to model training procedures or model exposure levels cannot be directly implemented. However, according to OWASP Top 10 for LLM Applications [21], most of them are recoverable as supply-chain

[12] https://genai.owasp.org/

Control	Applicability
Integrate ML specificities to awareness strategy and ensure all ML stakeholders are receiving it	A
Establish security process to maintain a good security level of the components of the ML application	NA
Use less easily transferable models	NA
Assess the exposure level of the model used	NA
Apply modifications to inputs	NA
Choose and define a more resilient model design	NA
Implement processes to maintain security levels of ML components over time	A
Assess the exposure level of the model used	NA
Enlarge the training dataset	NA
Apply an RBAC model, respecting the least privileged principle	A
Control all data used by the ML model	NA
Ensure reliable sources are used	NA
Integrate poisoning control after the ”model evaluation” phase	NA
Use methods to clean the training dataset from suspicious samples	NA
Ensure ML applications comply with identity management, authentication, and access control policies	A
Ensure that models are unbiased	NA
Ensure that the model is sufficiently resilient to the environment in which it will operate.	NA
Integrate ML applications into the overall cyber-resilience strategy	NA
Conduct a risk analysis of the ML application	A
Define and monitor indicators for proper functioning of the model	A (partly)
Ensure ML projects follow the global process for integrating security into projects	A
Ensure ML applications comply with third parties' security requirements	A (partly)
Assess the regulations and laws the ML application must comply with	A
Control all data used by the ML model	NA

Table 2: Vulnerability controls according to [8] and their applicability to AI Intelligent Coaching application.

controls (e.g., “Establish/maintain security process for ML components”), a subset is partially recoverable at the orchestration and retrieval layers (e.g, “Control all data used by the ML model” is applicable only for the Match to DB component), and only a residual core is genuinely un-closable without supplier attestation (e.g., cleaning/enlarging the training set and the model's intrinsic bias and robustness).

7. Conclusions

We applied the first four phases of FAITH_AI_TAF_V1 to an in-isolation assessment of the AI Intelligent Coaching application and its four AI assets. Insights from 17 workshops with 75 media professionals rank privacy as the highest business impact, followed by fairness and accuracy-and-robustness. Threat and vulnerability analyses show risk is strongly component-differentiated: the deterministic local Match to Debunked Claims module is least exposed, while the generative modules that ingest external content carry the highest recurring likelihoods. Based on OWASP LLM Top 10 (2025), the controls have been partitioned into three groups: those applicable to developer-owned components, those non-applicable as model-engineering controls but recoverable as supplier-governance controls, and a residual core—training integrity, intrinsic bias and robustness—closable only through supplier attestation. Future work will focus on systematic benchmarking of these controls against NIST AI RMF, OWASP LLM Top 10 (2025), and ISO/IEC 42001. Finally, study motivates future work on cascading effects, human vulnerabilities, and a dedicated control set for third-party-model AI.

Acknowledgments

The work presented here is funded by the European Union's Horizon FAITH project (Fostering Artificial Intelligence Trust for Humans towards the optimization of trustworthiness through large-scale pilots in critical domains), Grant agreement No: 101135932. The work presented here reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Gemini in order to: Grammar and spelling check. After using these tools, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] F. Theofanis, K. Kioskli, E. Seralidou, N. Polemi, V. Antoniou, A. Elbi, J. D. Meyere, S. Garsia, A. M. Correa, A. Carboni, S. Colantonio, S. Gravili, G. R. Leone, D. Moroni, M. A. Pascali, D. I. Fotiadis, M. Tsiknakis, N. Tachos, V. Pezoulas, D. Zaridis, G. Kaliatakis, A. Følstad, G. Mentzas, D. Apostolou, E. Anagnostopoulou, N. Grammatikos, E. Tsalapati, S. Modafferi, FAITH Methodological Framework and Requirements Analysis v1, Deliverable D2.1, Horizon Europe. Grant agreement No. 101135932, 2025.

[2] Z. Wang, Y. Pang, Y. Lin, Smart expert system: Large language models as text classifiers, arXiv preprint arXiv:2405.10523 (2024).

[3] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the fifth international workshop on natural language processing for social media, 2017, pp. 1–10.

[4] K. Kumar, T. Ashraf, O. Thawakar, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, P. H. Torr, F. S. Khan, S. Khan, Llm post-training: A deep dive into reasoning large language models, arXiv preprint arXiv:2502.21321 (2025).

[5] H. Nguyen, Z. He, S. A. Gandre, U. Pasupulety, S. K. Shivakumar, K. Lerman, Smoothing out hallucinations: Mitigating llm hallucination with smoothed knowledge distillation, arXiv preprint arXiv:2502.11306 (2025).

[6] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, H. Hajishirzi, When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, in: Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers), 2023, pp. 9802–9822.

[7] S. Ma, A. Salinas, J. Nyarko, P. Henderson, Breaking down bias: On the limits of generalizable pruning strategies, in: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 2025, pp. 2437–2450.

[8] B. Caroline, B. Christian, B. Stephan, B. Luis, D. Giuseppe, E. Damiani, H. Sven, L. Caroline, M. Jochen, D. C. Nguyen, et al., Securing machine learning algorithms, Athens, Greece: ENISA (2021).

[9] N. AI, Artificial intelligence risk management framework (AI RMF 1.0), URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai (2023) 100–1.

[10] X. Liu, P. Li, G. E. Suh, Y. Vorobeychik, Z. Mao, S. Jha, P. McDaniel, H. Sun, B. Li, C. Xiao, Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak llms, in: International Conference on Learning Representations, volume 2025, 2025, pp. 10313–10360.

[11] J. A. Leite, O. Razuvayevskaya, K. Bontcheva, C. Scarton, Llm-based adversarial persuasion attacks on fact-checking systems, arXiv preprint arXiv:2601.16890 (2026).

[12] B. Barakat, S. Jaf, Beyond traditional classifiers: Evaluating large language models for robust hate speech detection, Computation 13 (2025) 196.

[13] G. Yu, Z. Wang, Y. Huang, R. Zhong, Y. Zhong, Y. Wang, M. R. Lyu, Why does the llm stop computing: An empirical study of user-reported failures in open-source llms, arXiv preprint arXiv:2601.13655 (2026).
[14] A. E. Cinà, A. Demontis, B. Biggio, F. Roli, M. Pelillo, Energy-latency attacks via sponge poisoning, Information Sciences 702 (2025) 121905.
[15] A. A. Sami, G. Debnath, R. Dey, A. N. Chowdhury, Code poisoning through misleading comments: Jailbreaking large language models via contextual deception, in: 2025 28th International Conference on Computer and Information Technology (ICCIT), IEEE, 2025, pp. 3812–3817.
[16] W. Zou, R. Geng, B. Wang, J. Jia, PoisonedRAG: knowledge corruption attacks to retrieval-augmented generation of large language models, SEC '25, USENIX Association, USA, 2025.
[17] Y. Qu, Y. Liu, T. Geng, G. Deng, Y. Li, L. Y. Zhang, Y. Zhang, L. Ma, Supply-chain poisoning attacks against llm coding agent skill ecosystems, arXiv preprint arXiv:2604.03081 (2026).
[18] Y. Cui, S. Pan, Y. Liu, H. Zhang, C. Zuo, Vortexpia: Indirect prompt injection attack against llms for efficient extraction of user privacy, in: Findings of the Association for Computational Linguistics: EACL 2026, 2026, pp. 587–609.
[19] B. Metin, H. S. Karaca, F. Iradat, M. Wynn, Securing agentic ai with the nist cybersecurity framework 2.0, in: 2025 16th International Conference on Electrical and Electronics Engineering (ELECO), IEEE, 2025, pp. 1–5.
[20] O. G. S. Project, OWASP Top 10 for LLMs - GenAI Red Teaming Guide, 2025.
[21] OWASP Foundation, OWASP Top 10 for Large Language Model Applications (2025), Technical Report, Open Web Application Security Project (OWASP), 2024. URL: https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/, version 2025. Published by the OWASP GenAI Security Project.