PEQUOD: an Agentic AI Framework for Phishing Detection

Carlos Güemes-Palau*†, Pedro Casas† and Pere Barlet-Ros*

†*AIT Austrian Institute of Technology*, Vienna, Austria
**Barcelona Neural Networking Center, Universitat Politècnica de Catalunya*, Barcelona, Spain
carlos.guemes@upc.edu, pedro.casas@ait.ac.at, pere.barlet@upc.edu

***Abstract*—Phishing attacks remain one of the most prevalent forms of cybercrime, motivating the development of diverse detection approaches with different trade-offs in accuracy, robustness, interpretability, and computational cost. We introduce PEQUOD, an agentic AI framework for phishing detection that leverages Large Language Models (LLMs) to reason over complementary phishing-analysis techniques, including graph-based detection models, explainability methods, structural similarity retrieval, and external threat-intelligence sources through a unified MCP-based architecture. By combining heterogeneous sources of evidence and enabling adaptive analysis, PEQUOD provides more robust and interpretable phishing assessment under temporal drift. We present the architecture of PEQUOD and a preliminary evaluation on real-world phishing datasets, showing improved performance compared to the standalone models integrated within the framework.**

I. Introduction

The widespread impact of phishing attacks has motivated the development of diverse detection approaches. The simplest methods rely on blocklists of known malicious domains, such as Google Safe Browsing [1], OpenPhish [2], and PhishTank [3]. Other approaches analyze website content, HTML structure, and URL characteristics using heuristic rules [4], [5], Machine Learning (ML) and Deep Learning (DL) models [6], [7], or reference-based comparisons against legitimate brands and webpages [8]–[10]. However, these approaches present important limitations. Blocklists and heuristic methods are lightweight and efficient, but require continuous updates and often fail to detect newly emerging phishing websites. ML and DL models improve automation and flexibility, yet may struggle to generalize under distribution shifts and typically provide limited interpretability. Reference-based approaches often achieve higher accuracy, but at the cost of increased computational complexity and reliance on external resources and visual similarity analysis.

Large Language Model (LLM) agents have recently emerged as a powerful paradigm for building autonomous systems capable of reasoning, tool usage, and complex task orchestration. Beyond their strong reasoning capabilities, LLMs can interact through natural language, improving accessibility for non-technical users. In addition, protocols such as the Model Context Protocol (MCP) [11] enable these agents to integrate seamlessly with external tools, APIs, and services, significantly extending their functionality and adaptability.

We introduce PEQUOD, a framework for building LLM-based agents for phishing detection. PEQUOD leverages an MCP-based architecture that enables the agent to orchestrate multiple phishing-analysis tools, AI models, explainability methods, and external threat-intelligence services through a unified interface. The framework integrates recent graph-based models we have developed for phishing detection and explainability [7], [12], [13], together with public phishing blocklists and similarity-based retrieval over previously analyzed websites. By combining structural, contextual, and external threat-intelligence information, PEQUOD enables the agent to reason over complementary sources of evidence and obtain a more comprehensive view of a target webpage. PEQUOD incorporates a knowledge database that serves both as a cache for prior analyzes and as a long-term memory mechanism, allowing the agent to retrieve structurally similar websites and leverage historical knowledge during the analysis process.

The proposed approach serves two main objectives. First, by integrating multiple specialized tools, the LLM agent can aggregate and reconcile heterogeneous signals for phishing detection. Different tools capture complementary aspects of a website, such as visual appearance, structural properties of the HTML, and domain characteristics, enabling the agent to produce a more robust and comprehensive assessment. Furthermore, by leveraging its accumulated experience stored in the long-term memory database, the agent can adaptively weight the relevance of each tool based on past performance in similar scenarios.

Second, the use of LLM-based reasoning enables natural-language interaction and facilitates the interpretation of complex phishing analyses, improving accessibility for non-expert users. This capability is particularly relevant in phishing detection, where end users are often the primary targets of attacks and may lack the technical expertise required to interpret conventional cybersecurity tools and model outputs.

II. The PEQUOD Framework

Figure 1 presents an overview of the PEQUOD framework architecture. The client component connects to the MCP server and provides a natural-language interface between the user and the underlying LLM agent. The framework supports both

978-3-903176-83-6 ©2026 IFIP

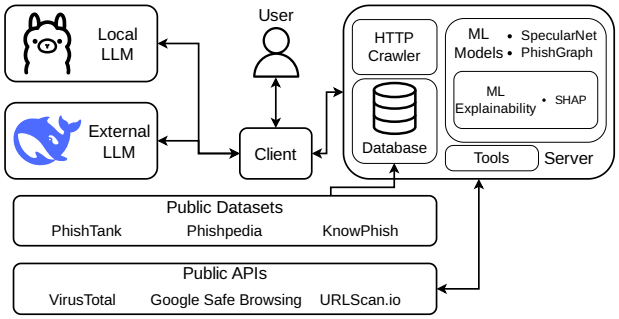

Fig. 1. Overview of PEQUOD framework. Comprised of a client and server process. The client offers an interface for the LLM model and connects to the server. The server process includes the offered tools and ML models, a knowledge database, and connections with other remote APIs.

locally deployed models (e.g., Ollama [14]) and externally hosted models accessed through APIs (e.g., DeepSeek [15]).

The server component is implemented as an MCP server that exposes a set of methods for phishing detection. It integrates core processing capabilities, including a lightweight web crawler for retrieving and parsing website HTML content, as well as a queue-based asynchronous task execution mechanism to handle computationally intensive analysis. The server further interfaces with external threat intelligence services via public APIs, such as Google Safe Browsing [1]. Incoming requests from the client are translated into structured API calls, enabling the LLM agent to seamlessly invoke these services through natural language queries while abstracting the underlying technical complexity.

The PEQUOD server integrates multiple AI models for phishing detection, including advanced graph-based approaches to analyze website structure and behavior. In particular, SpecularNet [7] leverages hierarchical graph representations of HTML structures together with domain-level features to perform reference-free phishing detection. PhishGraph [12] deploys a lightweight yet effective CatBoost-based architecture [16] that captures structural properties of webpages and their associated links. When invoked, the server retrieves the target webpage, extracts its HTML content and structural information, executes the corresponding phishing-detection models, and returns the resulting predictions to the LLM agent for further reasoning and analysis.

To enhance interpretability of the models' predictions, the framework integrates AI explainability techniques such as SHAP [17], along with extensions tailored to deep graph-based models, including DOM-SP-SHAP [13] for architectures like SpecularNet. These explainability methods can be invoked on demand to provide detailed insights into the factors driving the models' predictions. When queried, the server returns both the phishing-detection output and an explanation report generated using the corresponding explainability technique. The LLM agent can then incorporate this information into its reasoning process, enabling more transparent and context-aware responses to the user.

PEQUOD additionally includes a knowledge-base component that stores previously analyzed websites together with

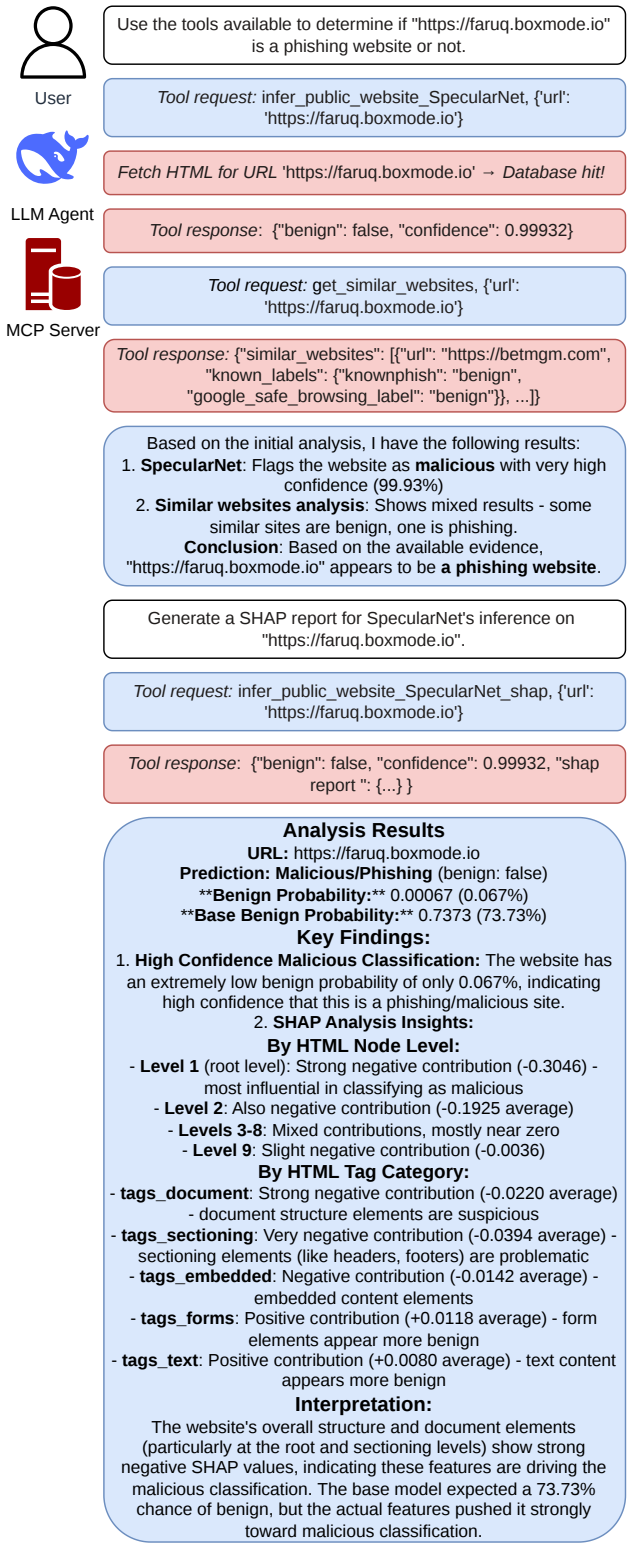

Fig. 2. Example interaction within the PEQUOD framework; the LLM agent coordinates phishing-detection models, explainability methods, and external sources to analyze a website through natural-language user queries.

their associated predictions and metadata. This knowledge base supports retrieval of structurally similar websites through HTML-based similarity indexing, enabling contextual reason-

ing over prior phishing analyses. This is particularly useful for long-running tasks and rate-limited APIs. Website similarity is computed from the HTML structure using path shingles [18]. This method represents a website through the set of nested HTML tag paths extracted from its DOM structure. For example, for the HTML '`<html><body><h1>Hello world</h1></body></html>`', the resulting set would be {[`html`], [`html, body`], [`html, body, h1`]}. Similarity between two websites consists of the Jaccard similarity between their two sets. The database is continuously populated during normal MCP server operation and can additionally be bootstrapped using publicly available phishing datasets, including PhishTank [3], Phishpedia [9], and KnowPhish [8].

Figure 2 illustrates an example interaction within PEQUOD, where the user communicates with the agent through natural-language queries, while the MCP server and the underlying LLM agent coordinate the execution of phishing-detection models, explainability methods, and external threat-intelligence sources.

III. Evaluation & Future Work

We present a preliminary evaluation of PEQUOD using an instance powered by DeepSeek V4 Flash [15]. The evaluation was conducted using the Holistic Agent Leaderboard (HAL) framework [19], a standardized infrastructure for reproducible evaluation of AI agents. The evaluated PEQUOD instance exposes the following functionalities:

❶ **Block-/Allow-lists:** Verification of whether a target website appears in the Google Safe Browsing API [1] or the Tranco Top 100K domains list (April 2026 snapshot) [20].

❷ **AI-Driven Detection:** Phishing analysis using the reference-free, graph-based detectors SpecularNet [7] and PhishGraph [12].

❸ **Explainability:** Generation of feature-level explanations by executing SpecularNet with DOM-SP-SHAP [13] and PhishGraph with SHAP [17], enabling interpretable analysis of the models' predictions.

❹ **Context Analysis:** Retrieval of structurally similar websites from the knowledge database using path shingles [18], together with their associated labels.

Both SpecularNet and PhishGraph are trained using the Phishpedia [9] dataset, which contains approximately 60k balanced phishing and benign websites collected in 2021. We further evaluate the models on two more recent and disjoint datasets: KnowPhish [8], comprising 10k balanced websites collected in 2023, and PhreshPhish [21], a large-scale real-world phishing dataset collected in 2025. For the latter, we randomly sample 500 balanced phishing and benign websites from its testing set.

Table I reports the accuracy, F1-score, precision, and recall of the evaluated approaches. Both models achieve strong performance on KnowPhish, reaching precision and recall values above 93%, indicating high phishing-detection capability with low false-positive rates. However, their effectiveness degrades significantly on the newer PhreshPhish dataset, highlighting the impact of temporal drift and the rapidly evolving nature of phishing websites. This motivates the use of PEQUOD as a reasoning and orchestration framework to analyze multiple complementary sources of information.

TABLE I
Performance comparison of phishing detection methods on the *KnowPhish* [8] and *PhreshPhish* [21] datasets.

Dataset	Model	Accuracy	F1-score	Precision	Recall
KnowPhish (2023)	SpecularNet	93.9%	93.9%	94.0%	93.9%
	PhishGraph	94.3%	94.2%	95.3%	93.1%
PhreshPhish (2025)	SpecularNet	52.7%	48.5%	54.3%	52.7%
	PhishGraph	75.6%	75.2%	77.3%	75.6%
	5-NN	63.0%	61.6%	65.3%	63.0%
	PEQUOD	**79.6%**	**79.6%**	**79.7%**	**79.6%**

The Phishpedia dataset is additionally included in the PEQUOD knowledge database and used as a reference base for similarity retrieval. In particular, we also evaluate a 5-Nearest Neighbor (5-NN) baseline on PhreshPhish, using the HTML shingle similarity metric to retrieve structurally similar websites to Phishpedia from the database. To ensure fairness and avoid information leakage, PEQUOD did not store evaluation results from PhreshPhish during the experiments.

PEQUOD consistently outperforms the individual phishing-detection models integrated within the framework. In particular, on the challenging PhreshPhish dataset, PEQUOD improves detection performance from 75% to nearly 80% as compared to the strongest standalone model, suggesting that reasoning over multiple complementary sources of evidence enables more robust phishing detection under temporal drift. Nevertheless, the results also highlight an important limitation: PEQUOD remains fundamentally dependent on the quality and generalization capabilities of its underlying detection models and data sources. While the framework can mitigate some weaknesses through multi-source reasoning and contextual analysis, it cannot fully compensate for severe degradation in the performance of the underlying phishing detectors.

These observations and preliminary results motivate several directions for further improvement. First, PEQUOD can be extended with additional phishing-analysis tools, including new threat-intelligence APIs, reference-based systems, and ML/DL detectors, enabling the agent to reason over a broader and more diverse set of complementary signals. Second, future versions of PEQUOD will incorporate uncertainty and prediction-confidence estimation techniques [22], allowing the agent to identify situations in which the underlying phishing-detection models may be unreliable and to adaptively request additional evidence or alternative analyzes. Third, the long-term memory and similarity-retrieval components of PEQUOD can be further enhanced through more advanced retrieval and ranking strategies, enabling improved contextual reasoning over previously analyzed websites and phishing campaigns. Finally, future work will consider larger-scale evaluations across additional datasets, temporally shifted scenarios, and real-world phishing campaigns, together with user-oriented studies assessing the interpretability, usability, and practical effectiveness of the framework for non-expert users.

ACKNOWLEDGMENTS

This work has been supported by the FWF Austrian Science Fund, Project reference I-6653, as well as by the Agencia Estatal de Investigacion, Project reference PCI2023-145974-2, as part of the EU CHIST-ERA-2022-SPiDDS-02 project GRAPHS4SEC. This work is also supported by the Catalan Institution for Research and Advanced Studies (ICREA Academia).

REFERENCES

[1] Google, "Google Safe Browsing." [Online]. Available: "https://safebrowsing.google.com/"
[2] "Openphish." [Online]. Available: "https://openphish.com/"
[3] "Phishtank." [Online]. Available: "https://www.phishtank.com/"
[4] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, "A novel approach for phishing detection using url-based heuristic," in *Int. Conf. Comput., Manage. Telecommun. (ComManTel)*, 2014, pp. 298–303.
[5] Y. Mourtaji, M. Bouhorma, D. Alghazzawi, G. Aldabbagh, and A. Alghamdi, "Hybrid rule-based solution for phishing url detection using convolutional neural network," *Wireless Commun. Mobile Comput.*, vol. 2021, no. 1, p. 8241104, 2021.
[6] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of ai-enabled phishing attacks detection techniques," *Telecommun. Syst.*, vol. 76, no. 1, pp. 139–154, 2021.
[7] T. Song, P. Casas, and M. Meo, "Phishing the Phishers with SpecularNet: Hierarchical Graph Autoencoding for Reference-Free Web Phishing Detection," *arXiv preprint, arXiv:2603.01874*, 2026.
[8] Y. Li, C. Huang, S. Deng, M. L. Lock, T. Cao, N. Oo, H. W. Lim, and B. Hooi, "KnowPhish: Large Language Models Meet Multimodal Knowledge Graphs for Enhancing Reference-Based Phishing Detection," in *USENIX Secur. Symp.*, 2024, pp. 793–810.
[9] Y. Lin, R. Liu, D. M. Divakaran, J. Y. Ng, Q. Z. Chan, Y. Lu, Y. Si, F. Zhang, and J. S. Dong, "Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages," in *USENIX Secur. Symp.*, 2021, pp. 3793–3810.
[10] R. Liu *et al.*, "Inferring Phishing Intention via Webpage Appearance and Dynamics: A Deep Vision Based Approach," in *USENIX Secur. Symp.*, 2022, pp. 1633–1650.
[11] Anthropic, "MCP - Model Context Protocol," 2024. [Online]. Available: "https://modelcontextprotocol.io/docs/"
[12] L. Torrealba, P. Casas, T. Song, J. Bustos-Jimenez, and I. Bachmann, "PHISHGRAPH – When HTML Graph Simplicity Outsmarts Deep and Reference-Based Phishing Detectors," in *Network Operations and Management Symposium (NOMS)*, 2026.
[13] C. Güemes-Palau, P. Casas, L. T. Aravena, T. Song, and P. Barlet-Ros, "DOM-SP-SHAP: Structure-Preserving Explainability for Graph-Based Phishing Detection," in *Proceedings of the 10th Network Traffic Measurement and Analysis Conference (TMA)*, 2026.
[14] "Ollama." [Online]. Available: "https://github.com/ollama/ollama"
[15] DeepSeek-AI, "DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence," 2026.
[16] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: Gradient Boosting with Categorical Features Support," 2018. [Online]. Available: https://arxiv.org/abs/1810.11363
[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[18] D. Buttler, "A short survey of document structure similarity algorithms," Lawrence Livermore National Laboratory (LLNL), Livermore, CA, 02 2004. [Online]. Available: https://www.osti.gov/biblio/15013935
[19] S. Kapoor *et al.*, "Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation," in *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. [Online]. Available: https://arxiv.org/pdf/2510.11977
[20] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation," in *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, ser. NDSS 2019, Feb. 2019.
[21] T. Dalton, H. Gowda, G. Rao, S. Pargi, A. H. Khodabakhshi, J. Rombs, S. Jou, and M. Marwah, "Phreshphish: A real-world, high-quality, large-scale phishing website dataset and benchmark," *arXiv preprint*, 2025. [Online]. Available: https://arxiv.org/abs/2507.10854
[22] J. Gawlikowski *et al.*, "A Survey of Uncertainty in Deep Neural Networks," *Artificial Intelligence Review*, vol. 56, no. 1, pp. 1513–1589, Oct 2023. [Online]. Available: https://doi.org/10.1007/s10462-023-10562-9