

Multimodal Neurosymbolic Agentic Systems for Cross-domain Intelligence Derivation

Pier Mattia Basciano and Salvatore Monteleone

Niccoló Cusano University, Rome, Italy

piermattia.basciano@unicusano.it

salvatore.monteleone@unicusano.it

Abstract: Modern organizations face an urgent paradox: AI systems grow increasingly capable yet progressively less interpretable - and increasingly dependent on external infrastructure that raises legitimate data sovereignty concerns. This paper addresses both tensions by presenting a multimodal neuro-symbolic agentic architecture that treats structured knowledge as the primary engine of intelligence derivation, evaluated across three high-stakes domains - cybersecurity, financial forecasting, and medical imaging - where explainability and privacy are functional requirements. The framework unifies sub-symbolic learning with expert-encoded formal logic through an agentic orchestration layer. All models run locally, ensuring sensitive organizational data never leaves the organization's own infrastructure. In the cybersecurity domain, LTN-induced rules govern malware classification across 24 families from the Maling dataset, with Redis storing Prolog-encoded logical rules alongside official Microsoft threat documentation. In the financial domain, expert-curated Prolog rules encode market patterns, risk indicators, and sector heuristics, with Redis retaining both formal logic and natural language counterparts. In the medical imaging domain, a CNN+LTN classifier trained on the COVID-19 Radiography Database produces both a diagnostic classification and the symbolic predicates justifying it. The agentic pipeline enforces two-stage validation: Qwen3:32b, as orchestrator, delegates image analysis to Qwen2.5-VL:7b, which verifies whether the radiological image is consistent with reported symptoms - only upon confirmed consistency is the neurosymbolic expert model invoked, otherwise the request is declined. All agents are evaluated via RAGAS across domain-specific query batteries: 24 cybersecurity queries (18 malware-specific, 6 cross-family disambiguation); 30 financial queries (15 highly logical, 15 general); and 30 medical queries (15 genuine COVID-19 cases, 9 valid non-COVID scenarios, 6 deliberately inconsistent inputs). For cybersecurity and financial use cases, an arena-style competition across six local models - DeepSeek-R1:32b, QwQ:32b, Gemma3:27b, Qwen3:32b, GPT-OSS:20b, and Phi4:14b - benchmarks the grounding effect of neuro-symbolic memory on response quality. Results demonstrate that symbolic grounding systematically drives model consensus toward correct classifications, positioning structured organizational knowledge as the active, privacy-preserving backbone through which agentic AI derives trustworthy intelligence - with direct implications for knowledge management in high-stakes learning organizations.

Keywords: Neurosymbolic AI, Agentic intelligence, Knowledge management, Explainability, Privacy-preserving AI

1. Introduction

Learning organizations thrive by cultivating cultures where individuals openly share information, reflect on practices, and collaboratively solve problems. Yet the rapid advancement of agentic AI introduces a fundamental challenge: can autonomous systems that identify patterns, forecast outcomes, and derive intelligence enhance organizational learning, or do they risk creating opaque dependencies that undermine the collaborative reasoning essential to knowledge-intensive institutions? This question becomes critical as organizations deploy AI across high-stakes domains - e.g., medical diagnosis, financial forecasting, and cybersecurity threat analysis - where decision errors carry severe consequences and where the provenance of every inference must remain auditable.

Nowadays, Large Language Models (LLMs) generate fluent intelligence but hallucinate unsupported facts, undermining trust, while Retrieval-Augmented Generation (RAG) systems improve factual grounding without guaranteeing logical consistency with domain principles. Neither approach enables expert teams to understand why a conclusion was reached, challenge underlying assumptions, or refine domain knowledge through collaborative reflection. The challenge is compounded in sensitive operational environments: organizations handling proprietary research data, patient records, or classified threat intelligence cannot route inference through external cloud services without exposing confidential information. Local deployment of AI systems addresses this constraint directly, keeping sensitive data within organizational boundaries while preserving full pipeline functionality. Furthermore, locally-hosted models introduce a natural scalability dimension: computational capacity determines model quality, so organizations can incrementally improve inference performance by scaling up hardware rather than accepting capability ceilings imposed by third-party providers.

Neuro-symbolic AI (NeSy) - integrating neural learning with symbolic reasoning - offers a technical solution aligned with these requirements. Unlike purely neural architectures that operate as black boxes, neuro-symbolic systems make reasoning explicit through symbolic constraints that expert teams can examine, audit, and refine. This transparency supports reflective practice: analysts can inspect which rules were activated, understand why

contradictions were flagged, and collaboratively improve the knowledge base (KB) as domain understanding evolves. However, existing neuro-symbolic approaches typically target single domains and modalities, failing to address the heterogeneous knowledge environments where learning organizations operate.

This paper presents three architecturally distinct neuro-symbolic pipelines that demonstrate how organizations can maintain explainability while enabling autonomous intelligence derivation across diverse high-stakes domains. Rather than proposing a monolithic architecture, our framework shows how neuro-symbolic principles can be adapted to different organizational knowledge contexts while preserving the transparency required for collaborative learning. Each pipeline orchestrates multimodal components - i.e., visual analysis, textual retrieval, logical inference - through agentic validation loops that render reasoning chains auditable. The technical foundations span three complementary methodologies: Logic Tensor Networks (LTNs) for learning symbolic predicates from visual data; expert-encoded Prolog predicates for formalizing validated domain principles; and RAG enhanced with symbolic constraint validation for grounding intelligence in documentary evidence. All components are deployed locally via self-hosted infrastructure, ensuring data sovereignty and hardware-bound scalability. Unified across all three approaches is the commitment to symbolic explainability: whether rules are learned from data or encoded by experts, they remain human-interpretable and logical. The contributions of this research are as follows:

- **Data-Driven Neurosymbolic Discovery:** the CNN+LTN approach can learn interpretable rules from labeled images, enabling expert teams to discover visual patterns while maintaining transparent symbolic reasoning chains for collaborative validation.
- **Expert Knowledge Formalization:** the Prolog-based approach, which encodes technical analysis principles validated over decades of data, demonstrates how established organizational expertise can be formalized into auditable symbolic rules generalizable across different domains.
- **Hybrid Collaborative Intelligence:** a combined approach integrating automated pattern detection with expert-encoded taxonomies, enabling technical knowledge sharing through standardized symbolic rule representations. This facilitates cross-team collaboration by establishing unified behavioral descriptions of analyzed elements, supporting replicable research methodologies, and consistent intelligence derivation across distributed organizational units.

The remainder of this paper is organized as follows: Section 2 reviews related work and background in neuro-symbolic AI architectures, RAG systems, and knowledge management for intelligent systems. Section 3 presents the methodology for each domain, detailing technical implementations across cybersecurity, finance, and medicine. Section 4 reports evaluation results, emphasizing accuracy, consistency, trustworthiness, and explainability, through an arena-based comparative analysis. Section 5 discusses cross-domain generalization, identified limitations, and implications for achieving organizational maturity in the age of AI. Section 6 concludes with future research directions for supporting authentic learning in AI-augmented organizations.

2. Background

The theoretical foundation for differentiable neuro-symbolic learning rests on LTNs (Badreddine et al. 2022), which embed first-order logic (FOL) into neural training through a fuzzy-based logic called Real Logic, enabling simultaneous optimisation of predictive accuracy and logical consistency. Odense and Garcez (2025) formally demonstrate that minimizing an LTN loss function is equivalent to maximising semantic fidelity with respect to a KB, directly underpinning the symbolic predicate extraction performed by our classifiers post-training. Taxonomically, Zhu and Sun (2025) map neuro-symbolic integration along three paradigms: Symbol for Neural, Neural for Symbol, and Hybrid. Our framework spans depending on the domain: the financial path operates as Symbol for Neural, while cybersecurity and medical paths are genuine Hybrids. Bougzime et al. (2025) further show that Ensemble NSAI architectures, where neural and symbolic components interact through a shared mediating layer, outperform all alternatives on consistency, explainability, and safety - matching precisely our multi-agent design. The normative dimension is formalised by Gaur and Sheth (2023) through the CREST framework (Consistency, Reliability, Explainability, Safety), whose notion of consistency - respecting symbolic constraints irrespective of query formulation - provides the theoretical motivation for the Symbolic Adherence (SA) metric introduced in our RAGAS evaluation.

In cybersecurity, Manthena et al. (2025) survey XAI techniques for malware detection, documenting the gap between opaque CNN classifiers and operational auditability requirements and positioning intrinsically interpretable methods above post-hoc attribution - e.g., SHAP, LIME. Our CNN+LTN addresses this gap directly by producing symbolic predicates as primary outputs. Younas et al. (2025) establish a strong performance

baseline of 97.98% on the Malimg dataset using CNN with transfer learning; our architecture extends this by augmenting classification with activated behavioural predicates. Anthony et al. (2024) apply Logic Explained Networks (LENs) to the EMBER dataset for malware detection with FOL explanations, confirming symbolic grounding viability while remaining limited to feature-vector inputs, unlike our visual representation approach. Hakim et al. (2025) in their survey on NeSy for cybersecurity identify LTNs as among the most promising techniques, while Malkawi and Alhajj (2026) identify explainability as the primary adoption barrier in vulnerability management and position neuro-symbolic architectures as the most promising response. Piplai et al. (2023) demonstrate knowledge-graph-guided reinforcement learning for malware detection that preserves analyst-interpretable reasoning traces - a complementary approach to our Prolog-grounded design. By contrast, Härer (2025) shows that purely neural multi-agent systems achieve high accuracy on cybersecurity tasks without symbolic constraints, illustrating precisely the logically inconsistent yet confident responses our SA score is designed to penalise.

In clinical AI, Lu et al. (2025) apply Logical Neural Networks (LNNs) with learnable thresholds to produce interpretable diagnostic rules from tabular data, establishing symbolic grounding as viable in healthcare decision support; our CNN+LTN specialist extends this to radiological image classification on the COVID-19 Radiography Database, for which CNN-only baselines exceed 93% accuracy based on the study of Constantinou et al. (2023), providing a performance reference for our symbolic pipeline. Magnini et al. (2022) introduce KINS, which injects Datalog rules into neural architectures prior to training; our approach inverts this direction: the LTN learns predicates from data and exports them as Prolog rules post-training, offering complementary trade-offs between prior-knowledge integration and data-driven rule discovery.

RAG systems underpin the documentary grounding layer in our cybersecurity and financial paths. Gao et al. (2024) survey Naive, Advanced, and Modular RAG paradigms, establishing that documentary retrieval substantially constrains confabulation in LLM pipelines. Saxena and Gaur (2026) extend this by introducing neuro-symbolic retrievers integrating symbolic structures into the retrieval phase itself, conceptually aligned with our design, where Prolog outputs and Redis retrievals are synthesised before LLM generation. For pipeline evaluation, Es et al. (2024) introduced RAGs: we adopt and extend it with a SA dimension, verifying output alignment with activated Prolog predicates.

Neuro-symbolic co-design has proven architecturally generalisable: Wickramarachchi et al. (2024) combine ontological knowledge graphs with sub-symbolic perception for autonomous driving, identifying emergent capabilities - e.g., explainability, semantic search, causal reasoning - that mirror the organisational knowledge management benefits our tri-domain framework targets. For evaluation methodology, Son et al. (2025) formalise Arena-Lite, demonstrating that tournament-based head-to-head comparisons yield more reliable LLM rankings than single-judge evaluation.

3. Methodology

Our neuro-symbolic framework is based on three architecturally distinct implementations, each employing different combinations of learned and expert knowledge while maintaining symbolic explainability. The neuro-symbolic components appear at multiple architectural layers, as shown in Figure 1, depending on domain requirements: as trained classifiers producing symbolic outputs, as expert-encoded rule bases consulted during reasoning, and as validation metrics ensuring logical consistency of generated intelligence.

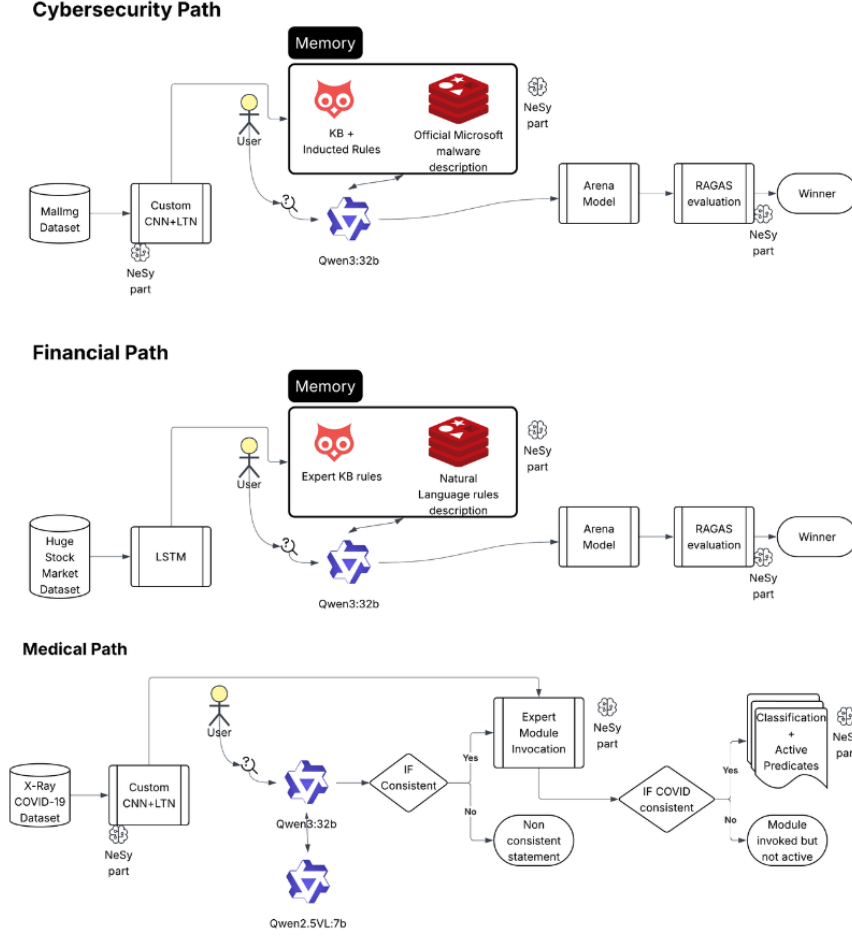


Figure 1: Pipeline across the three different use cases

The cybersecurity path combines a CNN+LTN classifier for malware analysis trained on Maling dataset with a hybrid Prolog KB containing both learned behavioral rules and expert-encoded attack taxonomies. RAG methodology provides the grounding documentation, with RAGAS evaluation metrics incorporating neuro-symbolic consistency checking to validate the threat behaviour against activated symbolic rules. The financial path employs expert-formalized Prolog predicates that encode more than 200 technical analysis principles, augmented by a Redis vector memory that stores natural-language rule explanations. Also this path utilizes RAGAS validation, ensuring generated forecasts satisfy activated Prolog constraints and preventing violations of fundamental market mechanics. The medical imaging path implements specialist-generalist orchestration based on an external expert module. A CNN+LTN expert module is trained on the COVID-19 Radiography Database, which produces symbolic diagnostic rules, which the general orchestrator (Qwen3:32b) interprets with secondary vision model (Qwen2.5-VL:7b) verification for consistency. Intelligence derivation relies entirely on learned symbolic rules and visual verification rather than documentary retrieval.

3.1 Cybersecurity Path: Malware Analysis

The cybersecurity intelligence derivation pipeline demonstrates how learned symbolic rules can support collaborative threat analysis in organizational security operations. The framework processes malware samples through five integrated stages. The pipeline begins with the Maling dataset, which contains binary visualizations of more than 9000 images from 24 malware families, represented as grayscale images. Preprocessing applies targeted data augmentation to family pairs that exhibit high confusion in the preliminary analysis. Then we implement a CNN+LTN training phase, producing a neuro-symbolic classifier that outputs both family predictions and activated symbolic rules describing behavioral characteristics. The model is trained on 29 predicates representing malware behaviors (e.g., *network_communication*, *registry_modification*, *elevates_privileges*) organized into ontologies. Post-training rule extraction generates rules, then stored in a Prolog KB. These inducted rules complement expert-encoded attack progressions from established threat taxonomies, creating a hybrid symbolic memory combining data-driven discoveries with domain expertise. The agentic orchestration layer employs Qwen3:32b as the central reasoning coordinator. When security analysts query the system - e.g.,

"The sample hides files on the filesystem, disables security software, uses obfuscation, and modifies system files to maintain persistence. Network activity is present, but no C2 server contact was confirmed. It does not steal credentials or replicate." -, the orchestrator retrieves relevant knowledge from two sources operating in parallel: (i) the Prolog KB previously described; (ii) a Redis-backed RAG system populated with Microsoft Security Threat Intelligence Center documentation describing known malware families, attack techniques, and mitigation strategies. The orchestrator synthesizes these inputs into natural language threat assessments, explaining both the classification and expected attack progression. To evaluate response quality beyond the orchestrator alone, an arena evaluation framework compares six locally-deployed models operating under identical neuro-symbolic constraints: Gemma2:7b, GPT-OSS:20b, Qwen3:32b, QwQ:32b-q4_K_M, DeepSeek-R1:32b, and Phi4:14b.

RAGAS evaluation employs a three-metric framework designed specifically for neurosymbolic intelligence validation, scoring each arena response through automated analysis by a judge model - i.e., Gemma2:7b. The evaluation framework computes three complementary scores with weighted aggregation:

- **F - Faithfulness** (40% weight): measures whether claims in the model's response are grounded in retrieved knowledge sources - both symbolic Prolog rules and documentary Redis. The judge model operates in two phases: first extracting atomic, verifiable claims from the candidate response, then systematically verifying each claim against the unified context assembled from Prolog reasoning and Redis retrieval. Claims contradicting retrieved evidence or introducing information absent from both knowledge sources reduce the faithfulness score proportionally. This metric quantifies the RAG contribution's effectiveness in constraining generation.
- **R - Answer Relevancy** (25% weight): evaluates whether the response addresses the specific query with pertinent, complete justification. The judge assesses four criteria: (i) Does the response provide a specific determination rather than vague alternatives - e.g., avoiding "*could be family X or Y*" without resolution for malware, or "*the market might rise or fall*" for financial forecasts - ? (ii) Is the conclusion motivated through domain-appropriate reasoning patterns described in the query? (iii) Does it avoid hedging without concluding? (iv) Is it coherent with the original query context rather than addressing a different specimen or market scenario? This metric ensures responses remain focused on the analytical task rather than producing tangential or incomplete assessments.
- **S - Symbolic Adherence** (35% weight): represents the core neuro-symbolic validation, measuring how strictly the model respected symbolic constraints from the Prolog reasoning layer. The judge verifies three critical alignment criteria: (i) Does the conclusion match the top-confidence candidate from Prolog symbolic reasoning? For cybersecurity, this means the classified malware family aligns with the highest-scoring Prolog candidate; for financial path, the forecast direction matches dominant activated predicates. (ii) If confidence percentages are mentioned, are they accurate within $\pm 5\%$ of Prolog-computed values? (iii) Are activated predicates from the symbolic layer actually referenced in the reasoning chain rather than ignored? Low SA indicates the model ignored neuro-symbolic constraints and reasoned autonomously.

The composite score aggregates all three metrics through weighted averaging ($0.40 \times F + 0.25 \times R + 0.35 \times S$), producing the final ranking. The highest-scoring model becomes the official arena winner.

3.2 Financial Path: Historical Data

Financial market intelligence requires balancing pattern recognition with fundamental principles validated across market regimes. The financial path formalizes expert knowledge through Prolog predicates encoding more than 200 logical rules organized into twelve technical analysis domains - e.g., *market trend classification*, *momentum indicators*, *volatility assessment*, *volume analysis*, *forecasting logic*. Each predicate implements fuzzy logic with a confidence score. A Redis vector memory stores natural language explanations for each rule, providing definitional descriptions. When Qwen3:32b retrieves Prolog rules, it simultaneously retrieves natural-language explanations from Redis, enabling forecasts that reference both logical activations and contextualized justifications. This expert-encoding approach reflects regulatory constraints requiring explainable decision logic and leverages decades of validated technical analysis principles, providing interpretable knowledge that analysts can audit and refine. LSTM temporal analysis on the Huge Stock Market Dataset informed query formulation by identifying which market patterns warranted investigation, guiding evaluation scenarios, and ensuring queries reflect statistically significant market phenomena analysts encounter in practice.

As for the Cybersecurity path, we perform an arena validation through six local models operating under identical neuro-symbolic constraints to evaluate the pipeline. RAGAS evaluation scores responses across the same three dimensions used for the above mentioned path.

3.3 Medical Path: Expert COVID-19 Module

The medical path demonstrates a distinct dimension of neuro-symbolic intelligence: the capacity for generalist orchestrators to invoke specialized expert modules trained for specific diagnostic tasks. While the cybersecurity and financial paths emphasize symbolic rule validation constraining LLM generation, the medical approach recognizes that certain knowledge-intensive tasks require dedicated neural-symbolic specialists operating beyond what general-purpose LLMs can reliably achieve. The pipeline operates by implementing a specialist-generalist orchestration where Qwen3:32b routes queries to appropriate expert modules based on clinical context. The agent performs initial triage, and if the query includes imagery, the orchestrator first invokes Qwen2.5-VL:7b for consistency verification. This approach prevents misrouted queries where symptoms contradict image findings - we will discuss it in Section 4. When consistency checks fail, the orchestrator declines to proceed, explaining the detected discrepancy rather than generating potentially misleading diagnostics that could be dangerous for the patient.

Upon passing verification, the orchestrator routes to the appropriate specialist from a registry of 30 available expert modules - e.g., *bone_fracture_expert*, *musculoskeletal_mri_expert*, *neurodegenerative_expert*. Critically, only one module is actually implemented - i.e., *covid_radiography_expert* - while the remaining 29 are placeholder stubs; the agent does not know which modules are active versus dormant. When routing to dormant modules, the system explicitly outputs that the specialist module invoked exists, but cannot be currently deployed, underlining the need of a consultation with human doctor. Meanwhile, when the active COVID-19 specialist is invoked, it returns structured neuro-symbolic output containing classification, confidence score, activated predicates with individual activation levels, differential diagnoses, and reasoning explanation. The orchestrator synthesizes this into a clinical assessment, preserving symbolic reasoning.

As for the cybersecurity path, we implement a CNN+LTN structure for the medical pre-training. We perform the analysis on the COVID-19 Radiography Database, containing more than 21000 chest X-rays images across four classes: *COVID-19*, *normal*, *lung opacity*, *viral pneumonia*.

4. Evaluation and Results

The evaluation framework assesses neuro-symbolic intelligence derivation across three methodological dimensions: classification accuracy for the medical path, symbolic rule compliance for financial and cybersecurity paths, and comparative quality assessment through multi-model arena competition. Query sets comprise domain-representative scenarios: 24 for the cybersecurity path, 30 for the financial path, and 30 for the medical path. All experiments were executed on local infrastructure equipped with a NVIDIA RTX 5090 GPU, an AMD Ryzen 9-9950X3D CPU, and 128GB of DDR5 RAM, enabling privacy-preserving deployment critical for organizational contexts that handle potentially sensitive data. The agentic orchestration pipelines were implemented in n8n (v2.7.5), a workflow automation platform for agent-based systems, through the implementation of dedicated code. We have selected a model based on architectural diversity. We chose three sequential generation models - i.e., Gemma2:7b by Google, Phi4:14b by Microsoft, and GPT-OSS:20b by OpenAI - two reasoning-specialized models - i.e., QwQ:32b-q4_K_M by Alibaba and DeepSeek-R1:32b - which utilize featuring extended chain-of-thought mechanisms, enabling a multi-step logical inference, and Qwen3:32b by Alibaba Cloud that serves also as the central orchestrator across all paths due to its balanced performance on both sequential generation and reasoning tasks. Finally, Qwen2.5-VL:7b is used for providing multimodal vision-language capabilities for medical consistency verification.

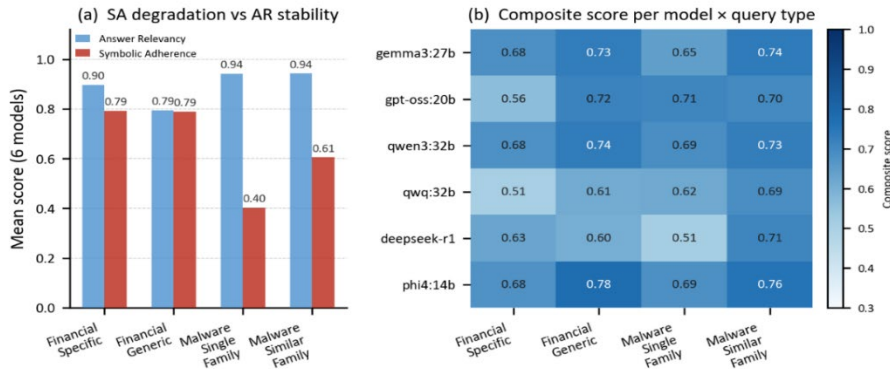


Figure 2: Arena Analysis

The cybersecurity arena evaluated 24 queries - 18 based on a single malware family and 6 based on a family with similar behaviour - measuring symbolic threat intelligence derivation from the Prolog KB augmented by a Redis NeSy-RAG layer. The underlying CNN+LTN classifier, trained on the Maling dataset across 24 malware families, achieves composite accuracy exceeding 99% with no evidence of overfitting. Qualitative inspection of arena voting patterns suggests that RAG activation reduces inter-model divergence on family-level queries, supporting the role of the Redis knowledge layer in grounding LLM responses within the structured symbolic space defined by the Prolog KB. Arena scores reveal a monotonic degradation of SA as a function of explicit KB coverage (Figure 2a): families with formal Prolog rules achieve a mean SA of 0.77, families covered only by rules induced during training drop to 0.53, and families entirely absent from the KB yield an adherence of 0.34 (Figure 3). This gradient confirms that lower SA scores on the malware path reflect intentional rule sparsity rather than model failure.

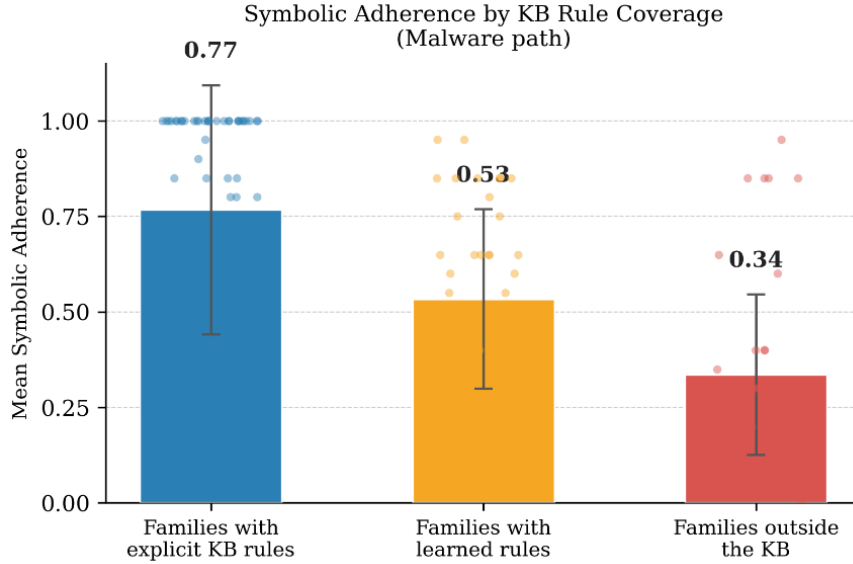


Figure 3: SA based on explicit and inducted KB

The financial path exhibits a qualitatively distinct profile, evaluated across 30 queries split equally between domain-specific questions - targeting the Prolog rules - and generic financial reasoning queries. Domain-specific queries yield SA of 0.793 against an AR of 0.898, demonstrating substantial symbolic adherence while maintaining fluent generation. Generic queries converge to near-parity between SA and AR, respectively 0.79 and 0.794, indicating that when queries do not require precise symbolic grounding, models appropriately leverage their pretrained knowledge. The SA/AR gap on domain-specific queries is significantly narrower than in cybersecurity contexts, suggesting expert-encoded Prolog rules effectively guide reasoning even in domains where LLMs possess baseline competence. This pattern validates the framework's design: symbolic grounding enhances rather than replaces model capabilities, ensuring logical consistency with technical analysis principles while preserving fluency essential for practical organizational deployment.

The medical pipeline was evaluated across 30 queries distributed over three scenario types: 15 COVID-consistent, 9 non-COVID-consistent, and 6 deliberately inconsistent cases (Table 1). The three-stage pipeline achieved correct end-to-end execution in 93.34% of COVID-consistent queries. The single failure constitutes a cascade error in which the orchestrator misrouted the query toward an unavailable specialist module, propagating the error across all three pipeline stages simultaneously. Non-COVID routing was perfect, and no false COVID diagnoses were produced on the inconsistent scenario set, confirming the pipeline's robustness against adversarial input inconsistency. The CNN+LTN classifier achieves a composite accuracy exceeding 90% with no sign of overfitting. Symbolic explainability is operationalised through LTN predicate activations: 12 of 29 radiological predicates activate on correctly classified COVID images (mean activation: 0.718), of which five reach near-certain activation (≥ 0.9) - green highlighted part in Table 1 - corresponding precisely to the established radiological semiology of COVID-19 pneumonia, confirming that the model has captured clinically meaningful features rather than statistical artefacts.

Table 1: Medical Path Results

Neurosymbolic Medical Pipeline — Evaluation Summary						
Scenario	N	Routing (a)	VL (b)	CNN+LTN classr. (c)	Outcome / Notes	Symbolic Explainability (LTN predicates, $\theta \geq 0.5$)
COVID consistent	15	14/15	14/15	14/15	1 routing error	12/29 active - mean 0.718 5 strong
Non-COVID consist.	9	9/9	9/9	—	specialist modules unavailable	—
Inconsistent	6	6/6	6/6	—	no false COVID diagnosis	—

LTN Top Activated Rule per Class — symbolic justification of classification			
COVID	Lung Opacity	Normal	Viral Pneum.
GGO $\wedge$ Periph. $\wedge$ Bilateral	Consolidation $\wedge$ Incr. opacity	Norm. parench. $\wedge$ $\neg$ Consolidation $\wedge$ $\neg$ GGO	Interst. thick. $\wedge$ Peribronchovasc. $\wedge$ Bilateral
$w = 2.21$	$w = 2.20$	$w = 2.76$	$w = 2.21$

(a) Qwen3:32b routing accuracy | (b) Qwen2.5-VL image-symptom consistency | (c) CNN+LTN classifications on successfully routed queries.
 Symb. expl.: active LTN predicates ($\theta = 0.5$); 5 strong ($\theta \geq 0.9$) = near-certain activations. w = learned rule weight (LTN training).

Across all query types, Figure 2(b) shows that phi4:14b achieves the highest composite arena score (mean: 0.725), demonstrating that parameter count is not the primary determinant of SA quality. Reasoning-oriented models - i.e., deepseek-r1:32b, qwq:32b - attain the lowest average scores (0.613 and 0.607, respectively), suggesting that extended chain-of-thought generation may introduce unverified inferences that reduce compliance with the symbolic KB. The most significant cross-domain finding is the persistent divergence between Answer Relevancy (AR) - uniformly high across all models and query types (0.79-0.94) - and SA, which degrades sharply in knowledge-intensive domains. This SA/AR gap constitutes the central empirical contribution of the arena evaluation: LLMs produce semantically coherent responses even when failing to satisfy formal symbolic constraints, underscoring the necessity of explicit neuro-symbolic grounding for trustworthy, auditable intelligence derivation in specialised organisational domains.

5. Generalization and Limitations

Nowadays, the neuro-symbolic reasoning operates strictly within predefined knowledge boundaries, with expressiveness bounded by domain-specific design choices. The cybersecurity path's predicates cannot discover novel behavioral concepts beyond expert-defined attributes. The financial path's rules can establish a technical analysis but cannot autonomously generate principles for unprecedented market dynamics. The medical path's specialist modules cover only explicitly trained conditions, unable to reason about unlisted pathologies. This knowledge-boundedness has two general implications: (i) interpretability guarantees apply only to pre-encoded concepts, and (ii) generalization to novel patterns requires manual knowledge shifting by updating the knowledge.

6. Conclusion and Future Work

This work demonstrates that neuro-symbolic agentic architectures can enhance the derivation of organizational intelligence while preserving the collaborative, reflective practices essential to learning organizations. Through three domain-specific implementations - learned visual rules via CNN+LTN for medical imaging, expert-encoded Prolog predicates for financial forecasting, and hybrid behavioral taxonomies for cybersecurity threat analysis - we show how organizations can leverage both data-driven discovery and accumulated expertise within unified symbolic validation frameworks. The persistent SA/AR divergence across all evaluated models constitutes the central empirical finding: LLMs generate semantically coherent responses (AR: 0.79-0.94) even when failing to satisfy formal symbolic constraints (SA: 0.34-0.79 depending on KB coverage). This gap underscores that fluency does not guarantee logical compliance, validating the necessity of explicit neuro-symbolic grounding for trustworthy intelligence in specialized domains.

Critically, the local deployment architecture enables privacy-preserving operation for sensitive use cases without relying on external APIs. This addresses a fundamental barrier to AI adoption in learning organizations: the need to maintain data sovereignty while leveraging advanced capabilities. The different symbolic KBs constitute shareable organizational assets that teams can audit, refine, and transfer across departments. When financial analysts propose new technical indicators or security teams encode emerging threat patterns, these updates propagate through the symbolic layer, enabling organizational learning rather than algorithmic dependence.

Future work should explore three directions aligned with learning organization evolution: (i) Automated predicate discovery through reinforcement learning or meta-learning to dynamically extend symbolic vocabularies as domain knowledge advances, reducing manual engineering overhead; (ii) Cross-domain knowledge transfer mechanisms enabling abstract reasoning patterns - e.g., trend reversal logic, anomaly detection heuristics - to inform rule initialization across related domains; (iii) Federated symbolic learning (Xing et al. 2024) where distributed organizational units share anonymized symbolic rules without exposing sensitive data, enabling collaborative intelligence across institutional boundaries while preserving privacy.

Ethics Declarations: This research did not require ethical approval as it did not involve human participants or sensitive data.

AI declarations: AI language models were used solely for language editing. All scientific content, analysis, and conclusions were developed by the authors, who take full responsibility for the work.

References

- Anthony, P., Giannini, F., Diligenti, M., Homola, M., Gori, M., Balogh, Š., & Mojžiš, J. (2024). Explainable Malware Detection with Tailored Logic Explained Networks. arXiv preprint. DOI:[10.48550/arXiv.2405.03009](https://doi.org/10.48550/arXiv.2405.03009)
- Badreddine, S., Garcez, A. d'A., Serafini, L., & Spranger, M. (2022). Logic Tensor Networks. Artificial Intelligence, 303, 103649. DOI:[10.1016/j.artint.2021.103649](https://doi.org/10.1016/j.artint.2021.103649)
- Bougzime, O., Jabbar, S., Cruz, C., & Demoly, F. (2025). Unlocking the Potential of Generative AI through Neuro-Symbolic Architectures: Benefits and Limitations. arXiv preprint. DOI:[10.48550/arXiv.2502.11269](https://doi.org/10.48550/arXiv.2502.11269)
- Constantinou, M., Exarchos, T., Vrahatis, A. G., & Vlamos, P. (2023). COVID-19 Classification on Chest X-ray Images Using Deep Learning Methods. International Journal of Environmental Research and Public Health, 20(3), 2035. DOI:[10.3390/ijerph20032035](https://doi.org/10.3390/ijerph20032035)
- Es, S., James, J., Anke, L. E., & Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 150–158). DOI:[10.18653/v1/2024.eacl-demo.16](https://doi.org/10.18653/v1/2024.eacl-demo.16)
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, M., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint. DOI:[10.48550/arXiv.2312.10997](https://doi.org/10.48550/arXiv.2312.10997)
- Gaur, M., & Sheth, A. (2024). Building trustworthy neurosymbolic AI systems: Consistency, reliability, explainability, and safety. AI Magazine, 45(1), 12–25. DOI:[10.1002/aaai.12149](https://doi.org/10.1002/aaai.12149)
- Hakim, S. B., Adil, M., Velasquez, A., Xu, S., & Song, H. H. (2025). Neuro-Symbolic AI for Cybersecurity: State of the Art, Challenges, and Opportunities. arXiv preprint. DOI:[10.48550/arXiv.2509.06921](https://doi.org/10.48550/arXiv.2509.06921)
- Härer, F. (2025). 'Specification and Evaluation of Multi-Agent LLM Systems - Prototype and Cybersecurity Applications', 2025 International Conference on Cybersecurity and AI-Based Systems (Cyber-AI), pp. 340–347. DOI: doi.org/10.48550/arXiv.2506.10467
- Lu, Q., Li, R., Sagheb, E., Wen, A., Wang, J., Wang, L., Fan, J. W., & Liu, H. (2025). Explainable Diagnosis Prediction through Neuro-Symbolic Integration. AMIA Joint Summits on Translational Science Proceedings, 2025, 332–341. PMID: 40502272 (PMCID: PMC12150699)
- Magnini, M., Ciatto, G., & Omicini, A. (2023). Knowledge injection of Datalog rules via Neural Network Structuring with KINS. Journal of Logic and Computation, 33(8), 1832–1850. DOI:[10.1093/logcom/exad037](https://doi.org/10.1093/logcom/exad037)
- Malkawi, M., & Alhajj, R. (2026). AI-Powered Vulnerability Detection and Patch Management in Cybersecurity: A Systematic Review of Techniques, Challenges, and Emerging Trends. Machine Learning and Knowledge Extraction, 8(1), 19. DOI:[10.3390/make8010019](https://doi.org/10.3390/make8010019)
- Manthena, H., Shajarian, S., Kimmell, J. C., Abdelsalam, M., Khorsandroo, S., & Gupta, M. (2025). Explainable Artificial Intelligence (XAI) for Malware Analysis: A Survey of Techniques, Applications, and Open Challenges. IEEE Access, 13, 61611–61640. DOI:[10.1109/ACCESS.2025.3555926](https://doi.org/10.1109/ACCESS.2025.3555926)
- Odense, S., & d'Avila Garcez, A. (2025). A semantic framework for neurosymbolic computation. Artificial Intelligence, 340, 104273. DOI:[10.1016/j.artint.2024.104273](https://doi.org/10.1016/j.artint.2024.104273)
- Piplai, A., Kotal, A., Mohseni, S., Gaur, M., Mittal, S., & Joshi, A. (2023). 'Knowledge-Enhanced Neurosymbolic Artificial Intelligence for Cybersecurity and Privacy', IEEE Internet Computing, 27(5), pp. 43–48. DOI: [10.1109/MIC.2023.3299435](https://doi.org/10.1109/MIC.2023.3299435)
- Saxena, Y., & Gaur, M. (2026). Neurosymbolic Retrievers for Retrieval-Augmented Generation. IEEE Intelligent Systems, 41(1), 96–104. DOI:[10.1109/MIS.2025.3642666](https://doi.org/10.1109/MIS.2025.3642666)

- Son, S., Oh, J.-M., Jin, H., Jang, C., Jeong, J., & Kim, K. (2025) 'Arena-lite: Efficient and Reliable Large Language Model Evaluation via Tournament-Based Direct Comparisons', in C. Christodoulopoulos et al. (eds) *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Suzhou, China: Association for Computational Linguistics, pp. 7057–7075. DOI:[10.18653/v1/2025.emnlp-main.360](https://doi.org/10.18653/v1/2025.emnlp-main.360)
- Wickramarachchi, R., Henson, C., & Sheth, A. (2024) 'Knowledge Graphs of Driving Scenes to Empower the Emerging Capabilities of Neurosymbolic AI', *IEEE Internet Computing*, 28(6), pp. 62–67. DOI:[10.1109/MIC.2024.3494972](https://doi.org/10.1109/MIC.2024.3494972).
- Xing, P., Lu, S. and Yu, H. (2024) 'Federated neuro-symbolic learning', in *Proceedings of the 41st International Conference on Machine Learning*. Vienna, Austria: JMLR.org (ICML'24).
- Younas, N., Riaz, S., Ali, S., Khan, R., Ali, F., & Kwak, D. (2025). Detecting malicious code variants using convolutional neural network (CNN) with transfer learning. *PeerJ Computer Science*, 11, e2727. DOI: [10.7717/peerj-cs.2727](https://doi.org/10.7717/peerj-cs.2727)
- Zhu, S., & Sun, S. (2024). A short Survey: Exploring knowledge graph-based neural-symbolic system from application perspective. arXiv preprint. DOI: [10.48550/arXiv.2405.03524](https://doi.org/10.48550/arXiv.2405.03524)