Explainable diagnostic intelligence: Neuro-symbolic approaches for medical imaging

Neuro symbolic AI: Making diagnostic predictions transparent, trustworthy and clinically aligned

Anindita Desarkar, PhD, CTO team, NTT DATA

Contents

1. Introduction

Diagnostic prediction is one of the most critical and complex challenges in healthcare, where the accurate and timely identification of diseases directly influences patient outcomes, treatment planning and clinical decision-making. Over the years, healthcare AI has evolved from traditional machine learning models to advanced deep learning architectures and, more recently, large language models (LLMs). While these technologies have demonstrated strong performance in multiple clinical use cases, they often operate as “black box” systems that have limited transparency and explainability.

In medical imaging and clinical diagnostics, explainability is essential. It plays a critical role in meeting regulatory requirements such as those set out in the Food and Drug Administration (FDA) guidelines and other emerging AI governance frameworks in healthcare. These regulations demand not only accuracy but also transparent reasoning pathways that can be audited and validated. For example, in radiology workflows, oncology screening and cardiology imaging, explainable AI ensures that diagnostic outputs can be traced back to interpretable features, supporting both clinical adoption and compliance. In addition, to establish trust, support clinical validation and ensure regulatory compliance, clinicians, radiologists and patients must understand the reasoning behind an AI-generated diagnosis.

This is where neuro-symbolic AI (NSAI) becomes highly relevant. NSAI combines the learning capabilities of neural networks with the logical reasoning and interpretability of symbolic AI to create more trustworthy and explainable diagnostic systems. Neural networks are highly effective in identifying complex patterns from medical images, signals and unstructured clinical data. However, they typically lack explicit reasoning and decision transparency. In contrast, symbolic AI provides structured reasoning, domain knowledge representation and interpretable decision logic, but it struggles with perception-based tasks and requires extensive manual rule engineering. By integrating these complementary approaches, neuro-symbolic AI enables healthcare systems to combine statistical learning with medical knowledge graphs, clinical rules and reasoning frameworks. This integration supports more accurate, explainable and clinically aligned diagnostic intelligence, helping healthcare organizations move toward transparent, trustworthy and human-centered AI-driven care.

2. The evolution of AI techniques

Figure 1 shows the evolution of various AI techniques along with their limitations. It also presents the motivation behind the neuro-symbolic approach.

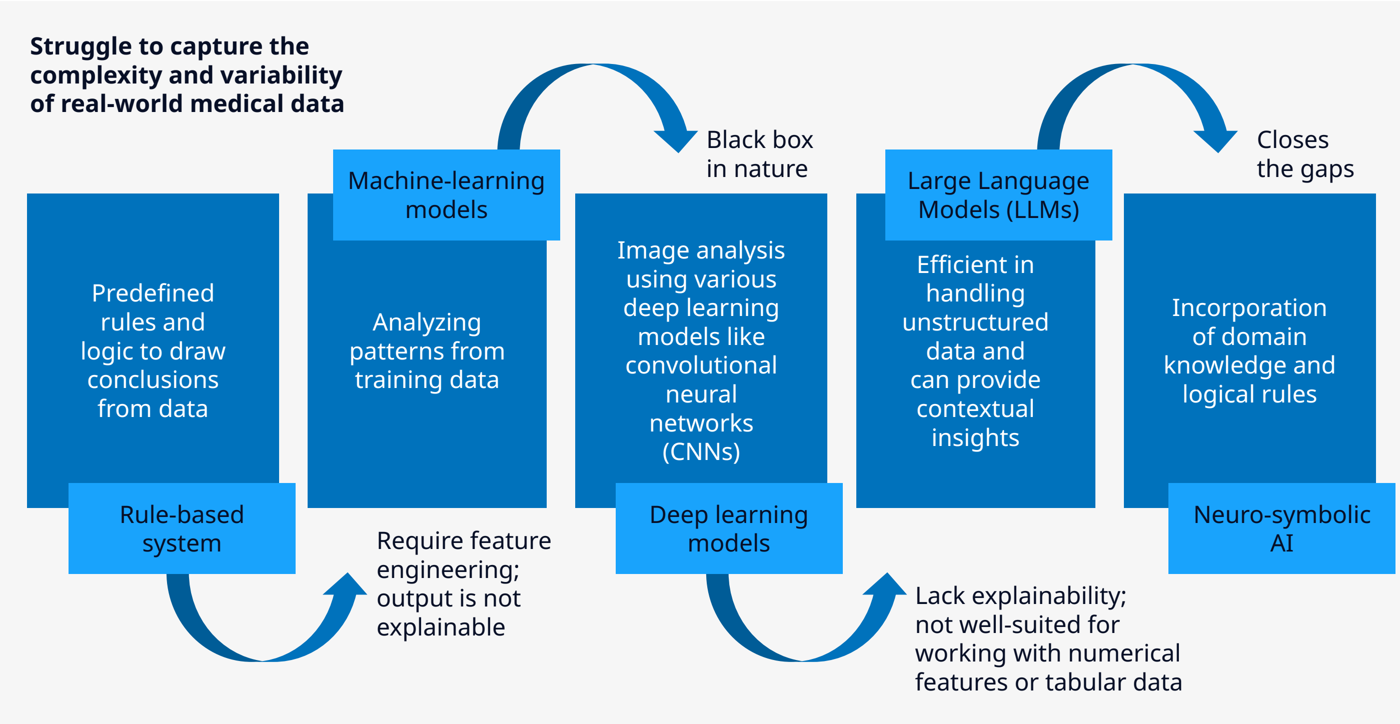


Figure 1: Evolution of AI techniques toward knowledge retrieval

Rule-based systems: Early diagnostic systems primarily relied on symbolic AI and predefined clinical rules to make decisions. These systems are highly interpretable because their reasoning process is transparent and easy to trace. However, rule-based approaches struggle to handle the complexity, variability and uncertainty present in real-world medical data.

Traditional machine learning models: Traditional machine learning techniques such as decision trees, support vector machines (SVMs) and logistic regression improved diagnostic prediction by learning patterns directly from historical data. These models provide better generalization and can process more complex datasets than rule-based systems. However, they depend heavily on manual feature engineering, which is time-consuming and may introduce domain bias. In addition, model interpretability decreases as algorithm complexity increases.

Deep learning models: Deep learning models, especially convolutional neural networks (CNNs), recurrent neural networks (RNNs) and transformers, have achieved high accuracy in medical imaging and clinical data analysis. CNNs are widely used for radiology and imaging tasks, while RNNs and transformers are effective for sequential healthcare data such as electronic health records (EHRs). Despite their strong predictive performance, these models often function as “black boxes” with limited explainability. They also require large volumes of labeled clinical data and high computational resources for training.

Large language models (LLMs): LLMs have introduced advanced capabilities for analyzing unstructured healthcare data such as clinical notes, discharge summaries and medical literature. They can understand medical context, generate human-like explanations and identify hidden relationships across large datasets. However, LLMs inherit many limitations of deep learning, including limited transparency, high computational cost and complex model behavior. They are also less effective in handling structured numerical and tabular healthcare data such as laboratory values and vital signs.

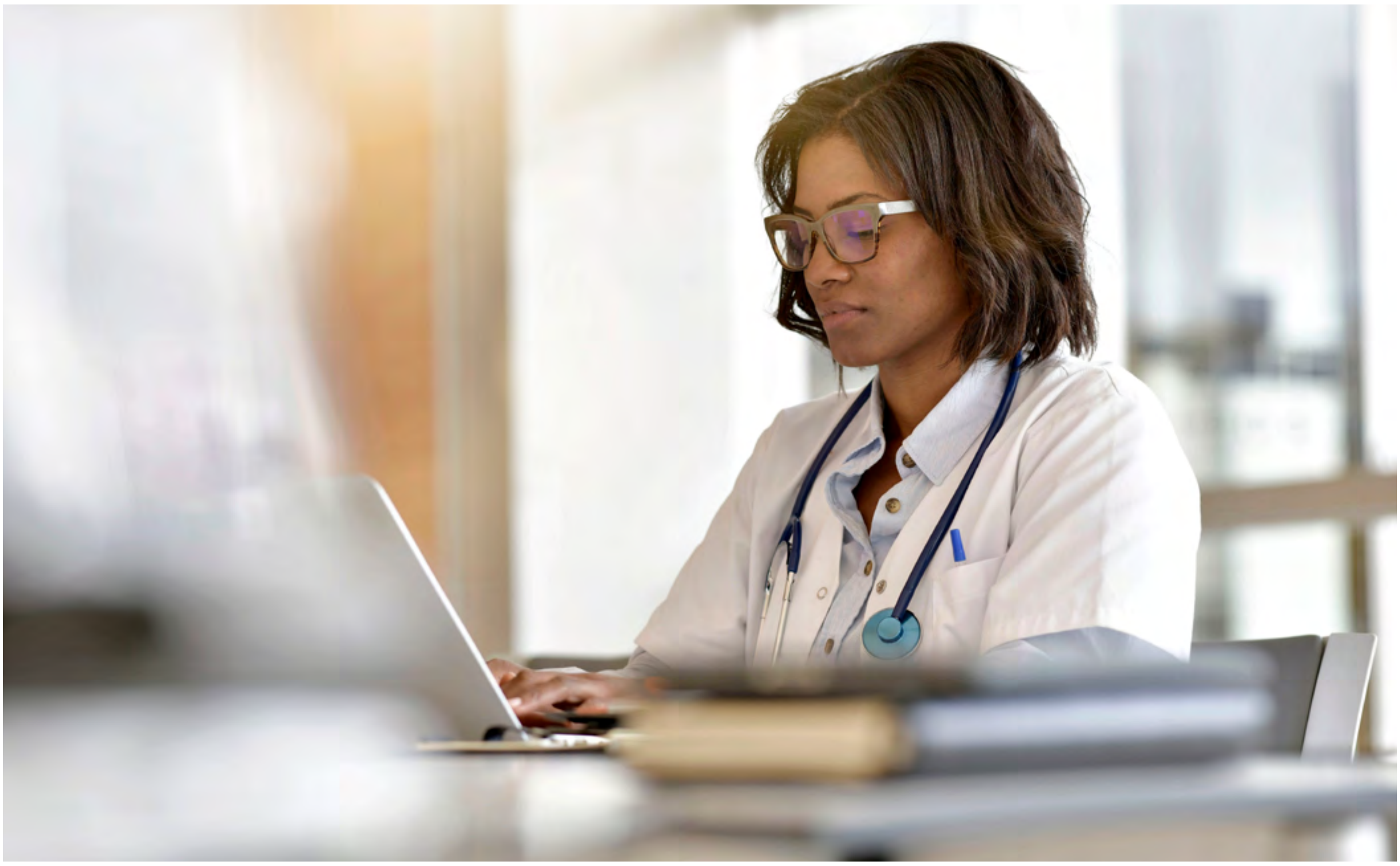
Neuro-symbolic AI (NSAI): Neuro-symbolic AI closes the above gaps. It combines the learning power of neural networks with the reasoning and explainability of symbolic AI.

Figure 2 presents the challenges addressed by this approach.

Traceability, explainability and hallucination	Scarcity of data	Robustness and trustworthiness	Energy inefficiency
Symbolic reasoning links outputs to explicit rules and knowledge graphs for auditability. Also, Neuro-symbolic Cycle is used during model training to make the output more deterministic.	Learning from fewer examples as integrated with rule-based system - Reducing reliance on massive datasets, making AI more sustainable.	Fewer chance of hallucination makes the application more robust and trustworthy.	Lesser dataset and lesser model training makes AI more sustainable. Saves energy and moves towards green computing.

Figure 2: Challenges addressed by the neuro-symbolic AI approach

While each of these approaches has advanced diagnostic intelligence in healthcare, neuro-symbolic AI should be viewed not as a replacement but as an augmentation of all prior techniques. By integrating neural perception with symbolic reasoning, NSAI enhances the strengths of rule-based systems, machine learning, deep learning and LLMs, while mitigating their limitations. Importantly, this augmentation directly supports regulatory compliance in life sciences: frameworks such as the Food and Drug Administration (FDA) in the United States and the Conformité Européenne (CE) marking in Europe increasingly emphasize transparency, auditability and explainability. NSAI provides traceable reasoning pathways that align with these requirements, ensuring that diagnostic outputs are both clinically trustworthy and regulation ready.



3. Challenges in medical imaging in existing techniques

3.1 Limited interpretability in model-driven diagnosis

Medical imaging is central to disease diagnosis and treatment planning in conditions such as diabetic retinopathy (DR), tumor detection and neurodegenerative disorders. While deep learning (DL) models, particularly convolutional neural networks (CNNs) and vision transformers (ViTs), have achieved remarkable predictive performance, their adoption in real-world clinical practice is hindered by their limited interpretability, as DL models are often black boxes and post hoc explainability methods such as Grad-CAM and SHAP remain heuristic, static and disconnected from clinical reasoning.

The lack of interpretability in model-driven diagnosis can be broken down into three dimensions:

- **Clinical disconnect:** Deep learning outputs often fail to align with the reasoning process clinicians use, making adoption difficult.
- **Heuristic limitations:** Post hoc methods such as Grad-CAM and SHAP provide static visualizations that remain disconnected from medical logic, reducing their reliability in real-world practice.
- **Regulatory barriers:** Healthcare regulations such as FDA and CE standards require transparent, auditable reasoning pathways, which black-box models cannot provide.

Together, these factors highlight why interpretability is not just a technical preference but a clinical and regulatory necessity in medical imaging AI.

3.2 Long-tailed distribution and imbalanced representation learning

Clinically critical pathologies are inherently rare and exhibit high phenotypic heterogeneity. Traditional deep learning architectures rely on the assumption of uniform data distribution and optimize for overall accuracy; consequently, they develop a systemic algorithmic bias toward the majority classes.

This skewed optimization suppresses the gradients of underrepresented categories, leading to poor feature extraction and a failure to capture the nuanced, low-prevalence visual and clinical biomarkers necessary to identify minority disease classes.

Figure 3 presents these various issues step by step.

Core machine learning problem: Long-tailed distribution	In medical imaging, data naturally follows a long-tailed distribution. A few conditions (like "normal/healthy") make up the massive head of the curve, while hundreds of rare diseases form a long tail, with very few images per disease.
Gradient suppression: Machine learning models fail	Deep learning models learn by calculating errors (gradients) and adjusting their weight. • If 99% of your dataset is "healthy" and 1% is a "rare tumor," the model can achieve 99% accuracy simply by guessing "healthy" every time. • The signals from the 1% "rare tumor" are mathematically drowned out (suppressed) by the massive volume of signals from the "healthy" data.
Clinical complication: Phenotypic heterogeneity	"Rare" is only half the problem. The other half is heterogeneity, meaning the rare disease looks completely different from patient to patient depending on the stage of the disease, the patient's anatomy, and imaging technology. The model is forced to try to learn a highly complex, variable pattern from only a handful of examples.
Engineering consequence: Poor feature representation	Because the model does not see enough diverse examples of the minority class, it fails to build robust latent space representations (the internal mathematical definitions of what a disease looks like). Instead of learning the true biological markers of the disease, it might just memorize irrelevant background noise in those few images, leading to poor generalization on new patients.

Figure 3: Challenges in interpreting medical data

3.3 Cross-domain generalization

Models trained on one institution's data frequently fail on data from other centers because of variations in acquisition protocols, imaging devices or patient demographics.

3.4 Integration complexity

Medical imaging modalities such as MRI, CT, X ray, ultrasound and PET generate vast, heterogeneous datasets. Variability in acquisition protocols, noise, limited labeled datasets and lack of interpretability in deep learning models create challenges in decoding medical data.

4. Architectural paradigms in neuro-symbolic AI

Neuro-symbolic architecture has two separate building blocks: a neural component and a logical component. These can be integrated in various ways, for example, sequentially, in parallel or some nested way. Various architectural patterns exist for each integration style, and each pattern is suitable for a specific use case.

Figure 4 presents various patterns or architectures for neuro-symbolic AI.

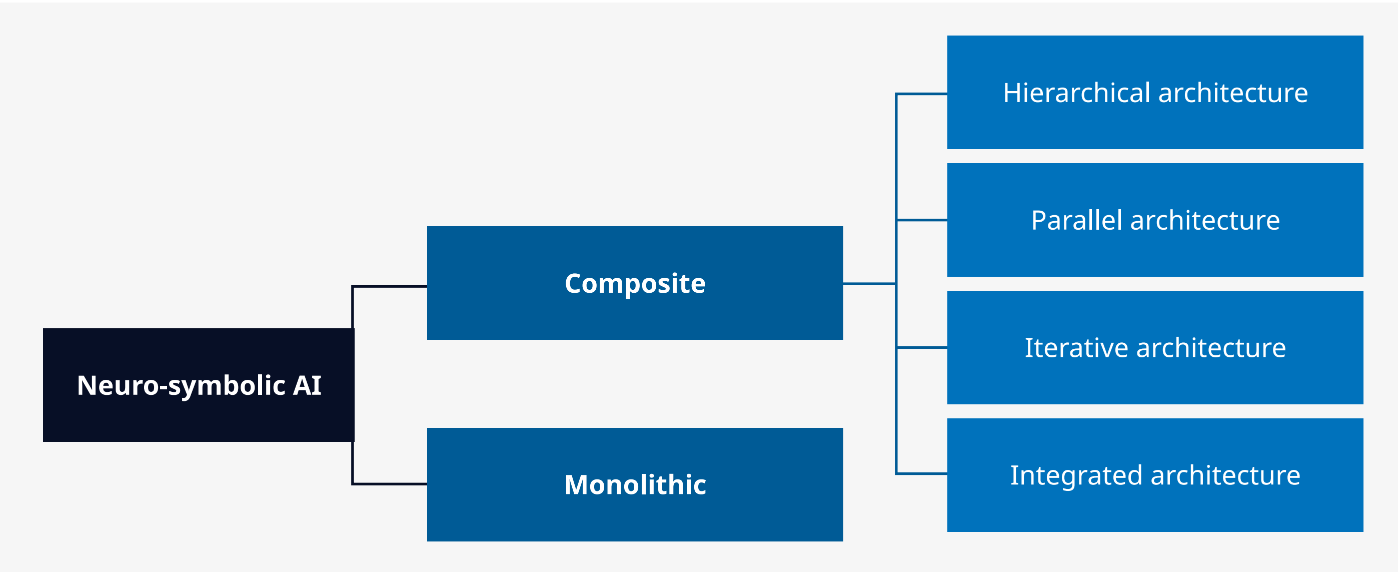


Figure 4: Architectural patterns in neuro-symbolic AI

A monolithic or composite pattern is based on how neural and symbolic components are integrated at a software and system level. In monolithic neuro-symbolic architecture, the neural and symbolic capabilities are fused into a single, unified mathematical engine or computational graph. The boundary between "perception" and "logic" is blurred because the symbolic rules are compiled directly into the neural network architecture. Key characteristics include a single training loop and differentiable logic. With a single training loop, the entire system is optimized from end to end using a single loss function. Gradient descent updates both the perceptual features and the logical representations simultaneously. A Logical Neural Network (LNN) is one such implementation.

On the other hand, a composite neuro-symbolic architecture treats the neural networks and symbolic engines as distinct, autonomous components or agents that interact through well-defined software interfaces. This is a highly modular, decoupled design. Key characteristics include decoupled runtimes and orchestration driven architecture. With decoupled runtimes, the neural component (for example, a large language model or an object detection model) runs on its inference engine, while the symbolic component (for example, a knowledge graph or failure mode and effects analysis (FMEA) logic tree) runs on a separate deterministic execution engine. Another important feature is orchestration, where a centralized controller or agentic workflow routes inputs through the neural perception layer, passes the extracted concepts to the symbolic reasoning engine, and aggregates the results based on the chosen topology: hierarchical, parallel or iterative.

In summary, in monolithic frameworks, the neural and symbolic components are fused into a single computational graph, meaning they cannot be isolated because rules are compiled directly into the network and optimized via a single training loop. In contrast, in composite frameworks, components are distinct and autonomous, running on decoupled engines where an orchestrating layer routes intermediate data between the neural perception and symbolic reasoning layers to compile the final output.

A brief overview of each approach is given below.

4.1 Hierarchical architecture

Neural then symbolic (or vice versa): Hierarchical architectures organize neural and symbolic components in layered structures, and information flows through distinct processing stages.

In these systems, a neural approach is used to predict the initial outcome and symbolic logic acts as a filter, ensuring that the AI model adheres to predefined constraints. This layered approach enables the clear separation of concerns and facilitates debugging and maintenance. For example, a deep learning model may detect a condition in medical diagnosis, while symbolic reasoning ensures recommendations align with medical guidelines. This is the easiest approach and therefore the most common and widely used.

Figure 5 presents the approach.

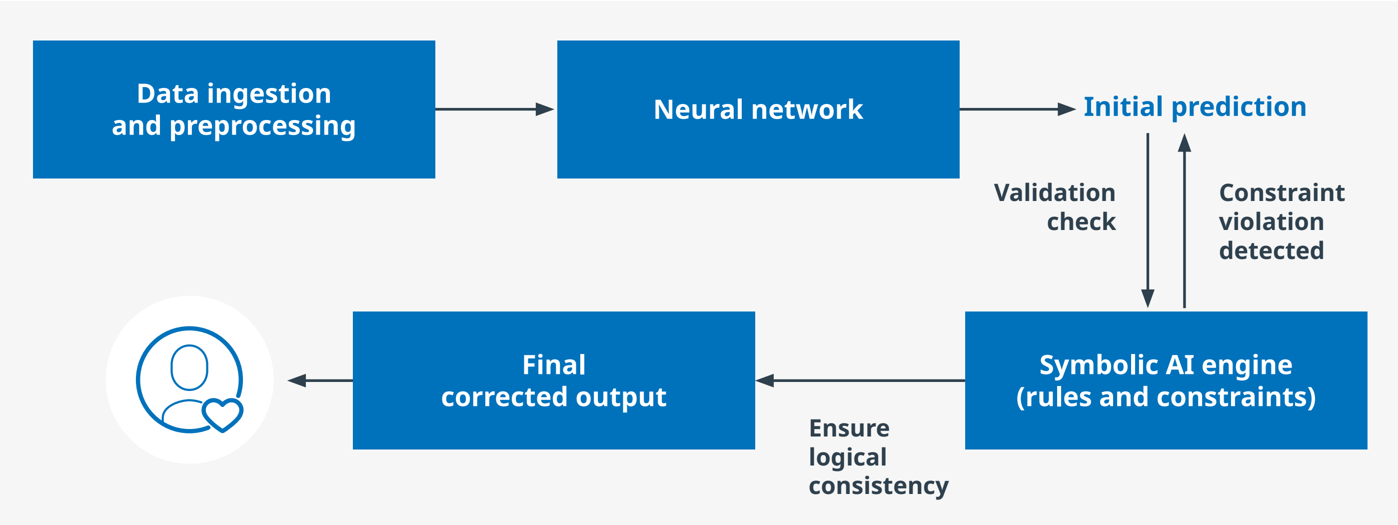


Figure 5: Neuro-symbolic design: Hierarchical architecture[1]

[1] and Prabhakar, AV. Neuro-Symbolic AI for Multimodal Reasoning: Foundations, Advances, and Emerging Applications. July 27, 2025.

4.2 Parallel architecture

Parallel architecture operates neural and symbolic components concurrently, combining their outputs through various fusion mechanisms. This approach enables real-time processing by leveraging the parallel nature of neural and symbolic computation. Fusion strategies include weighted voting schemes, attention-based combination mechanisms and probabilistic integration methods. These architectures are particularly effective for time-critical applications where rapid response is essential, such as autonomous vehicle control or real-time human–robot collaboration.[2]

Concordia is an example of a parallel neuro symbolic framework that runs a neural network and a probabilistic logic model side by side, letting both predict the same output.[3] A gating network learns how much to trust the neural prediction versus the logic based prediction for each input (see following figure 6).

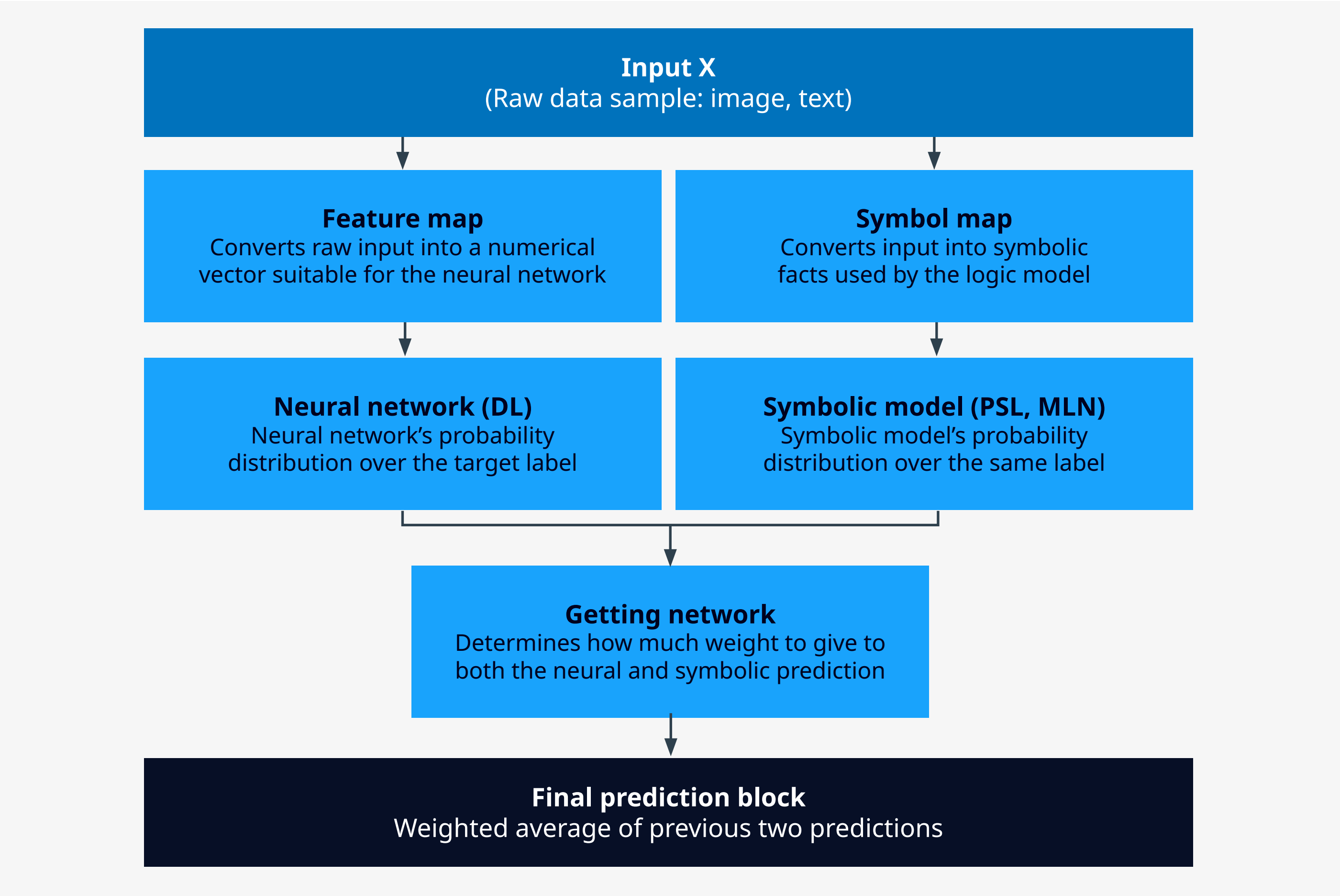


Figure 6: Concordia: A parallel neuro symbolic architecture

4.3 Nested architecture

Integrated architecture achieves deeper fusion by embedding symbolic reasoning within neural network structures or incorporating neural processing within symbolic frameworks.
As an example, in symbolic [neuro] systems, neural networks serve as subroutines within a predominantly symbolic framework — for instance, using neural nets to evaluate game positions within a symbolic search algorithm like AlphaGo's Monte Carlo tree search. Conversely, neuro [symbolic] systems embed symbolic reasoning within neural architectures — such as when a neural network encodes a maze into latent representations, then decodes symbolic action sequences like "forward, turn left, turn right" to solve it.[4]

[2] Prabhakar, AV. Neuro-Symbolic AI for Multimodal Reasoning: Foundations, Advances, and Emerging Applications. July 27, 2025.

[3] Feldstein, J., Dilkas, P., Belle, V., & Tsamoura, E. (2024). Mapping the neuro-symbolic AI landscape by architectures: A handbook on augmenting deep learning through symbolic reasoning. arXiv preprint arXiv:2410.22077.

[4] and Prabhakar, AV. Neuro-Symbolic AI for Multimodal Reasoning: Foundations, Advances, and Emerging Applications. July 27, 2025.

4.4 Iterative architecture

In cooperative architecture, neural and symbolic components work as interconnected partners, collaborating iteratively.

Neural networks process unstructured data like images, converting them into symbolic representations. The symbolic component evaluates and refines these representations, providing structured feedback that guides the neural network's updates. For autonomous vehicles, this means a neural network detects traffic signs, while a symbolic reasoning engine verifies detections against contextual rules — flagging inconsistencies like a stop sign in the middle of a highway and prompting the neural network to reevaluate until reaching consistent, high-confidence decisions.[5]

One such example is presented in Section 6.3, showing how the iterative approach improves end performance.

Figure 7 presents the schematic diagram of this architecture.

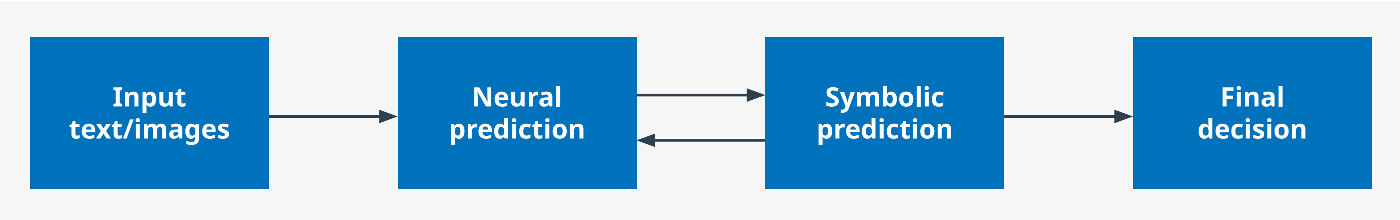


Figure 7: Iterative architecture

[5] Prabhakar, AV. Neuro-Symbolic AI for Multimodal Reasoning: Foundations, Advances, and Emerging Applications. July 27, 2025.

5. Frameworks in neuro-symbolic AI: the technical toolkit

5.1 Logic Tensor Networks (LTNs)

Logic Tensor Networks (LTN) are a neuro-symbolic framework through which querying, learning and reasoning are enabled using both rich data and abstract world knowledge. Within LTNs, a fully differentiable logical language known as real logic is introduced, in which elements of a first order logic signature are grounded in data by means of neural computational graphs and first order fuzzy logic semantics.[6]

5.2 Logical Neural Networks (LNNs)

A Logical Neural Network (LNN) is a form of recurrent neural network with a 1-to-1 correspondence to a set of logical formulas in any of various systems of weighted, real-valued logic, in which evaluation performs logical inference [14][10].

5.3 DeepProbLog

DeepProbLog is a system that makes us write logic programs (like Prolog), but inside those programs we can call neural networks as if they were normal logical predicates. It turns this whole combination into one big differentiable model, so we can train: the neural networks, the probabilistic facts and the logic program's parameters all together using gradient descent [9]. Following figure 8 presents the same.

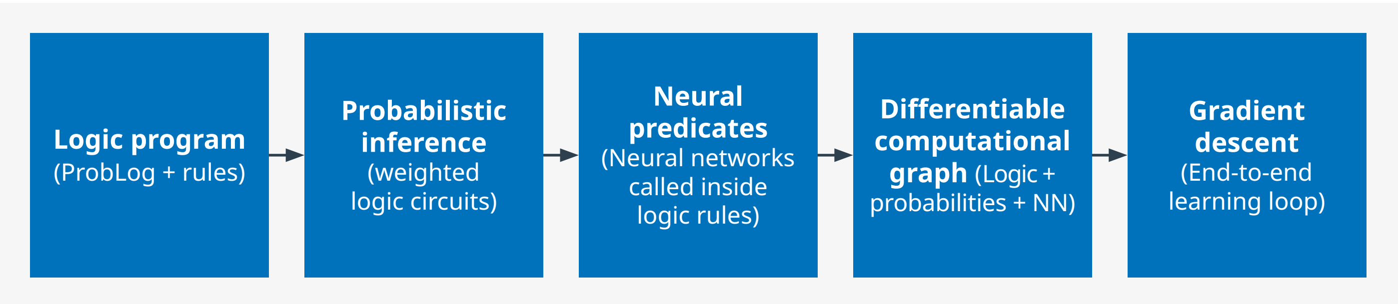


Figure 8: DeepProbLog framework: Working principle

5.4 Scallop

Scallop is a neuro-symbolic programming language that combines deep learning with logical reasoning so developers can build AI systems that learn efficiently and reason reliably. It has three main features: a flexible symbolic data model based on relations; a Datalog style logic language that supports recursion, aggregation, and negation; and a differentiable reasoning engine.[7]

[6] Riegel, R., Gray, A., Luus, F., Khan, N., Makondo, N., Akhalwaya, I. Y., ... & Srivastava, S. (2020). Logical neural networks. arXiv preprint arXiv:2006.13155.

[7] Neuro-Symbolic AI with AllegroGraph

5.5 AllegroGraph

AllegroGraph is a high performance knowledge graph database that combines symbolic reasoning with neural network based learning to support powerful neurosymbolic AI applications. It integrates knowledge graphs, a VectorStore, and deep LLM Integration (see figure 9).[8]

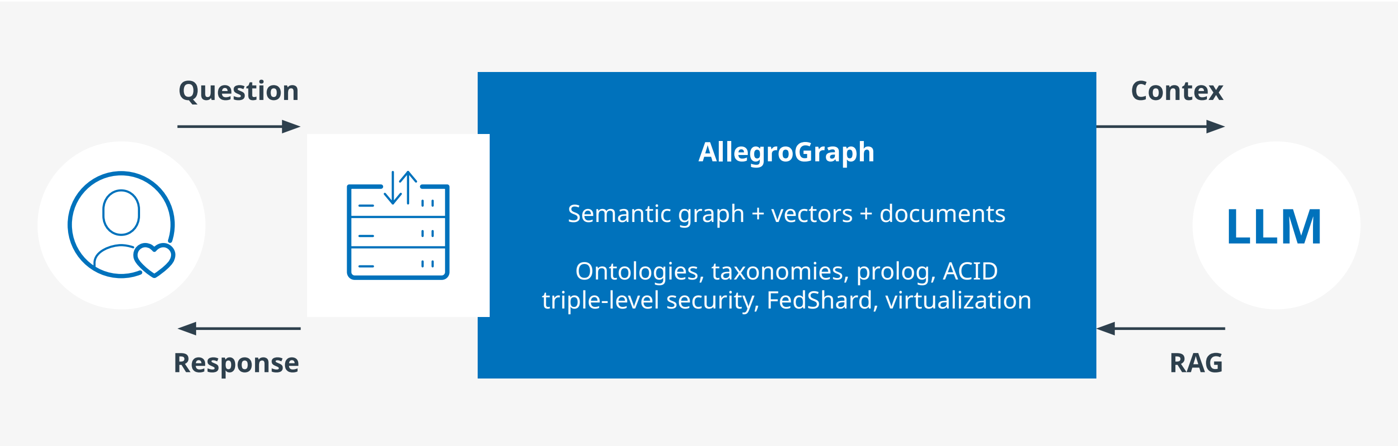


Figure 9: AllegroGraph: A high-performance graph database

Table 1 presents the comparison of the above approaches in the context of various real-life applications.

Framework	Core idea	Applications
Logic Tensor Networks (LTNs)	Differentiable logical language (real logic) combining neural graphs with fuzzy logic semantics	Medical imaging pattern recognition with explainable rules; clinical decision support
Logical Neural Networks (LNNs)	Neural units mapped directly to logical formulas for inference	Transparent diagnostic reasoning; interpretable risk scoring in healthcare
DeepProbLog	Logic programming (Prolog) extended with neural predicates; trained jointly via gradient descent	Probabilistic diagnosis; multimodal medical reasoning combining text and images
Scallop	Neuro symbolic programming language with relational data model, Datalog logic and differentiable reasoning	Clinical natural language processing (NLP) pipelines; reasoning over EHRs with recursion and aggregation
AllegroGraph	High performance knowledge graph database integrating symbolic reasoning, vector stores and LLMs	Knowledge graph driven diagnostics; biomedical research knowledge retrieval

Table 1: Comparison of various neuro-symbolic frameworks

[8] Badreddine, S., Garcez, A. D. A., Serafini, L., & Spranger, M. (2022). Logic tensor networks. Artificial Intelligence, 303, 103649.

6. Exploring neuro-symbolic frameworks in medical imaging

6.1 NeuroSymAD: A neuro-symbolic framework for Alzheimer's disease diagnosis

Alzheimer's disease (AD) diagnosis is challenging because it requires combining brain imaging with clinical information to make reliable decisions. Traditional deep learning models can analyze MRI scans well, but they often behave like black boxes and struggle to incorporate essential clinical factors such as biomarkers, patient history and demographic data.

Frameworks like NeuroSymAD merge computer vision with LLM based reasoning.[9] The neural network handles MRI feature extraction, and the LLM transforms established medical knowledge into logical rules. The symbolic module then reasons through a patient's biomarkers, age and medical history to classify whether mild cognitive impairment (MCI) is likely to progress to Alzheimer's disease (see figure 10).

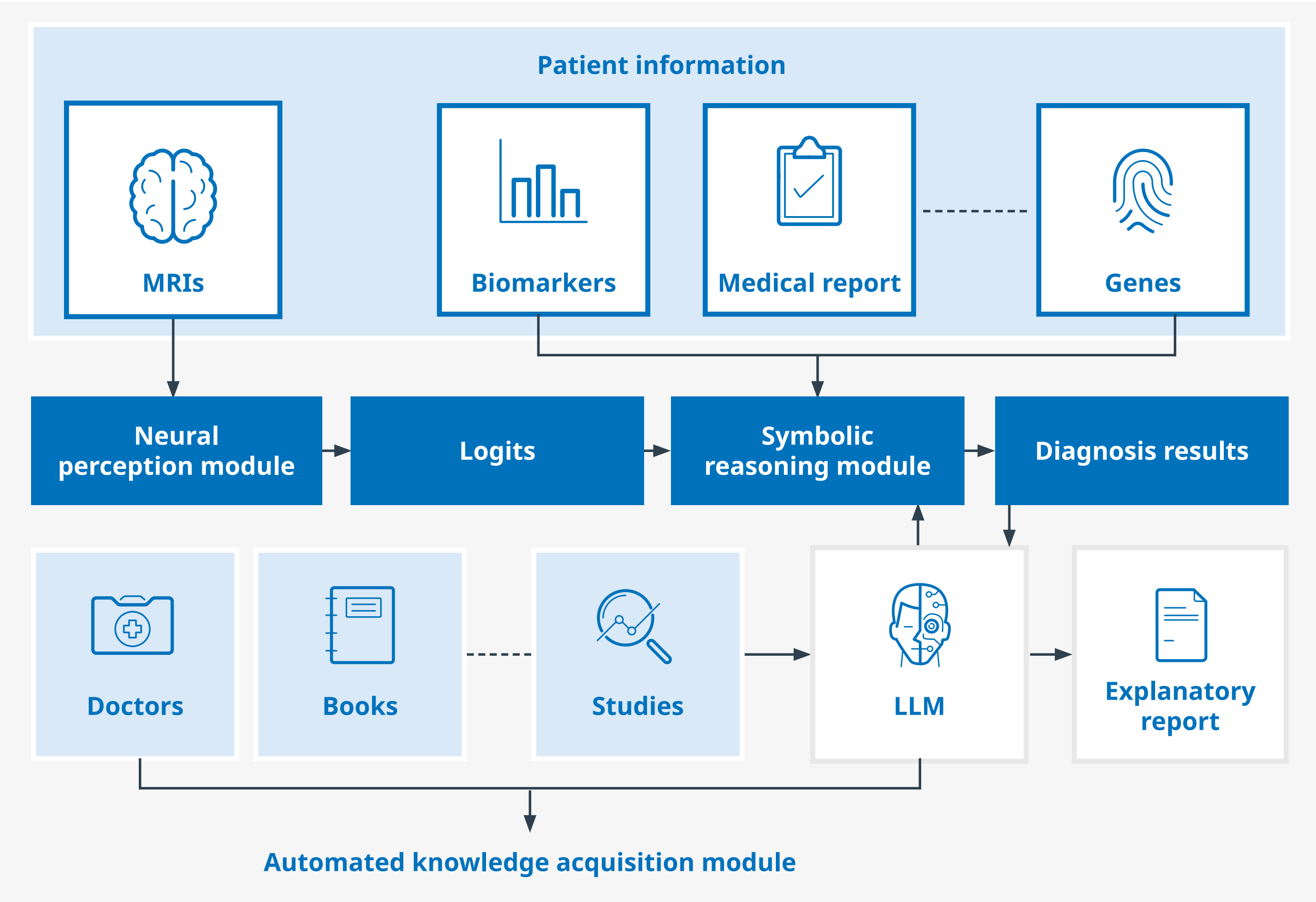


Figure 10: A Neuro-symbolic framework for Alzheimer's disease diagnosis

[9] Prenosil, G. A., Weitzel, T. K., Bello, S. C., Mingels, C., Manzini, G., Meier, L. P., ... & Afshar-Oromieh, A. (2025). Neuro-symbolic AI for auditable cognitive information extraction from medical reports. Communications Medicine, 5(1), 491.

Figure 11 shows an example of NeuroSymAD's diagnostic process and the generated explanatory report, demonstrating how the symbolic reasoning module corrects the misclassification of neural network.

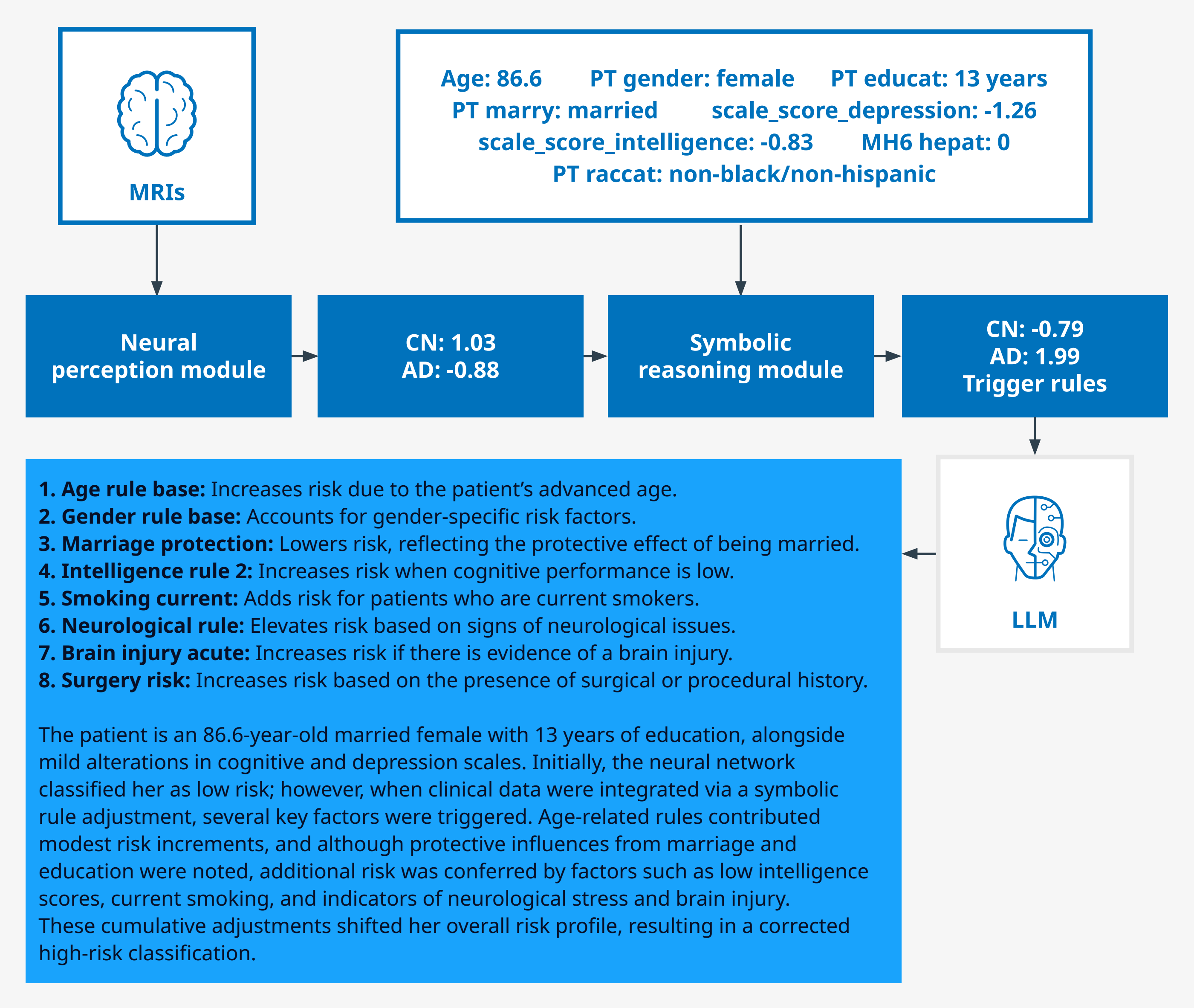


Figure 11: Example of NeuroSymAD's diagnostic process

With traditional deep learning, the problem is that models act like black boxes — they can read MRI scans but fail to use important clinical details like biomarkers, history or demographics. Neuro symbolic AI solves this by combining neural networks with symbolic rules so that both images and patient information are used together for better diagnosis. NeuroSymAD follows a hierarchical neuro-symbolic paradigm: neural perception is used for MRI features and symbolic reasoning is used for clinical rules. This improves accuracy and makes the system more trustworthy for doctors.

6.2 Neuro-symbolic AI for auditable cognitive information extraction in radiology

A unified semantic neuro symbolic NLP pipeline is designed so that each component compensates for the limitations of the others.[10] In a 2025 study, 206 anonymized PET/CT prostate cancer reports were processed through this pipeline. The reports were first segmented and checked to ensure proper anonymization. GPT 4 was then used to extract 26 predefined clinical parameters from the free text sections, and these extracted elements were converted into structured semantic facts. A locally run rule based expert system (Plato 3) validated and corrected these facts using medical ontologies, resolved contradictions, and performed higher level clinical inference such as distinguishing primary tumor staging from recurrence, identifying pathology and interpreting prostate-specific antigen (PSA) context.

Figure 12 presents the approach.[11]

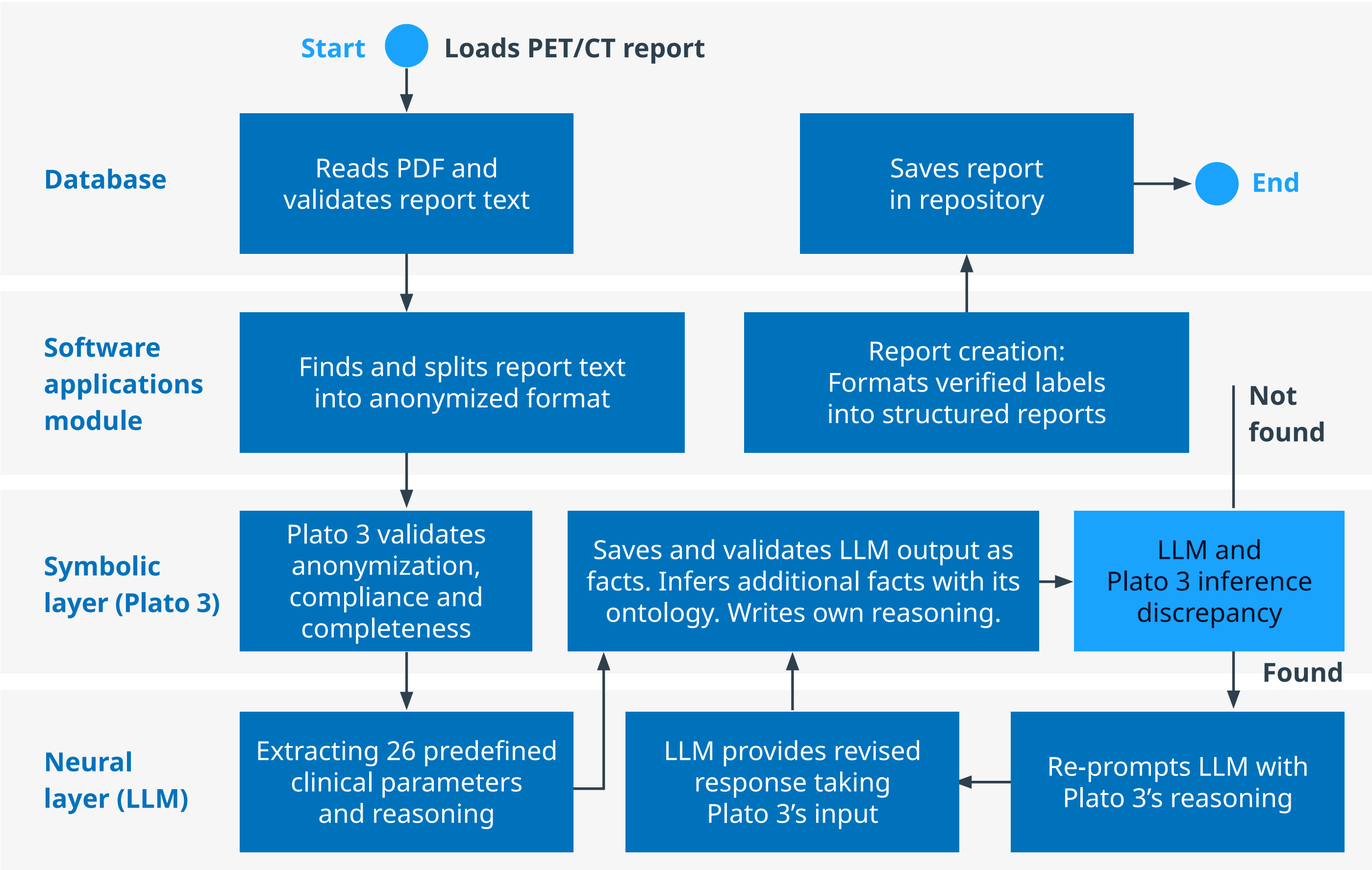


Figure 12: Framework for auditable cognitive information extraction from medical reports

With conventional NLP approaches, LLMs can extract parameters but often make errors, lack auditability and struggle with compliance. Neuro symbolic AI addresses this by combining GPT 4's neural extraction with a symbolic expert system (Plato 3), which validates facts, resolves contradictions and ensures regulatory alignment. Here, an iterative architecture is used: Neural perception is used for text extraction, symbolic reasoning is used for validation, and there is iterative feedback between the two. This architecture is chosen because it balances accuracy with transparency, making the pipeline auditable and trustworthy for clinical use.

[10] García-Barragán, Á., Sakor, A., Vidal, M. E., Menasalvas, E., Gonzalez, J. C. S., Provencio, M., & Robles, V. (2025). NSSC: a neuro-symbolic AI system for enhancing accuracy of named entity recognition and linking from oncologic clinical notes. Medical & biological engineering & computing, 63(3), 749-772.

[11] Prenosil, G. A., Weitzel, T. K., Bello, S. C., Mingels, C., Manzini, G., Meier, L. P., ... & Afshar-Oromieh, A. (2025). Neuro-symbolic AI for auditable cognitive information extraction from medical reports. Communications Medicine, 5(1), 491.

6.3 NSSC: Neuro-symbolic system for enhancing the accuracy of named entity recognition and linking from oncologic clinical notes

Clinical oncology notes contain rich but unstructured information. Extracting entities (for example, tumor types, treatments, biomarkers) and linking them to standardized vocabularies is essential for cancer research, clinical decision support, patient stratification and treatment optimization. However, a study in a Spanish oncology setting found that clinical notes lacked specialized NLP tools, and existing systems relied heavily on string similarity and faiedl to use context, causing incorrect mappings and no full pipeline exists for oncology text.[12]

The Neuro-Symbolic System for Cancer (NSSC) is designed to provide structure and semantics to unstructured short medical notes, addressing the critical need for normalization of clinical notes within specific medical domains, such as oncology. The primary motivation for developing NSSC is the lack of robust tools and methods capable of effectively transforming these unstructured texts into structured data, which is crucial for advancing clinical research, decision-making and patient care.[13]

The NSSC approach formulates the problem as an optimization task, with the goal of maximizing the accuracy of medical entity recognition and linkage while minimizing the associated costs. NSSC uses a hybrid AI model that integrates three distinct paradigms: deep learning (a language model trained with annotated medical entities), symbolic reasoning (symbolic background knowledge) and generative AI (LLM to solve the optimization problem), each of which contributes uniquely to NLP tasks.

By orchestrating components based on these three paradigms, NSSC effectively recognizes medical entities and links them to standardized medical terminologies within controlled vocabularies, such as the Unified Medical Language System (UMLS). Furthermore, the modular design of NSSC ensures its generalizability across different medical domains beyond oncology.[14] This approach is presented in figure 13.

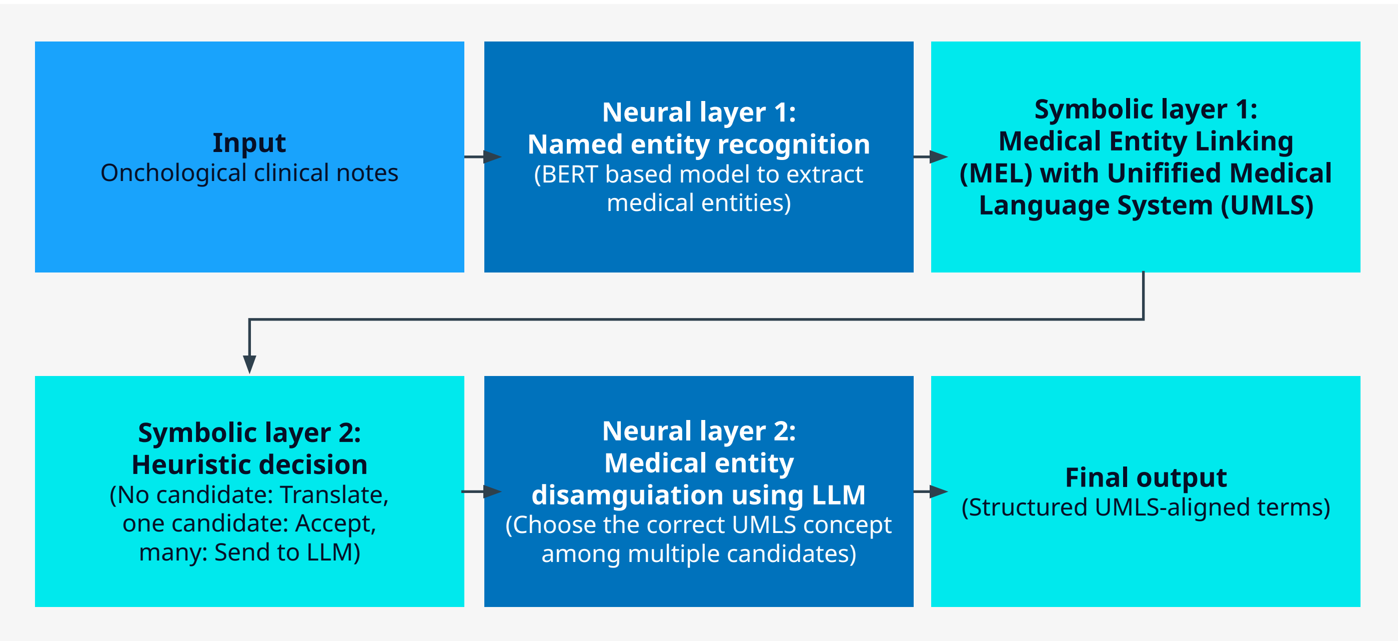


Figure 13: Methodology for Neuro-Symbolic System for Cancer (NSSC)

Conventional AI approaches often struggle with entity recognition and linking because they rely too heavily on string similarity, ignoring the broader context and leading to frequent incorrect mappings. This challenge is especially pronounced in Spanish oncology notes, where specialized tools are limited. Neuro symbolic AI overcomes these shortcomings by combining neural models for entity extraction, symbolic knowledge bases for linking and validation, and LLMs for disambiguation. This integrated approach delivers greater accuracy, contextual awareness and semantic consistency in clinical text analysis.

[12, 13, 14] Feldstein, J., Jurčius, M., & Tsamoura, E. (2023, July). Parallel neurosymbolic integration with Concordia. In International Conference on Machine Learning (pp. 9870-9885). PMLR.

7. An agentic system for recommending the best neuro-symbolic architecture and design based on use case

Based on previous studies, hierarchical architecture remains the most widely adopted due to its simplicity. However, alternative designs like parallel, integrated and iterative can offer superior performance depending on the specific requirements of the use case. So, the key challenge, therefore, lies in selecting the most appropriate architecture given the use-case requirements and constraints.

The figure 14 introduces a novel agentic system capable of recommending the optimal neuro symbolic design.

The system consists of four primary agents:

- Use case analyzer
- Architecture selection
- Design generator
- Evaluation

It will work in a sequential manner as each task is dependent on the previous one.

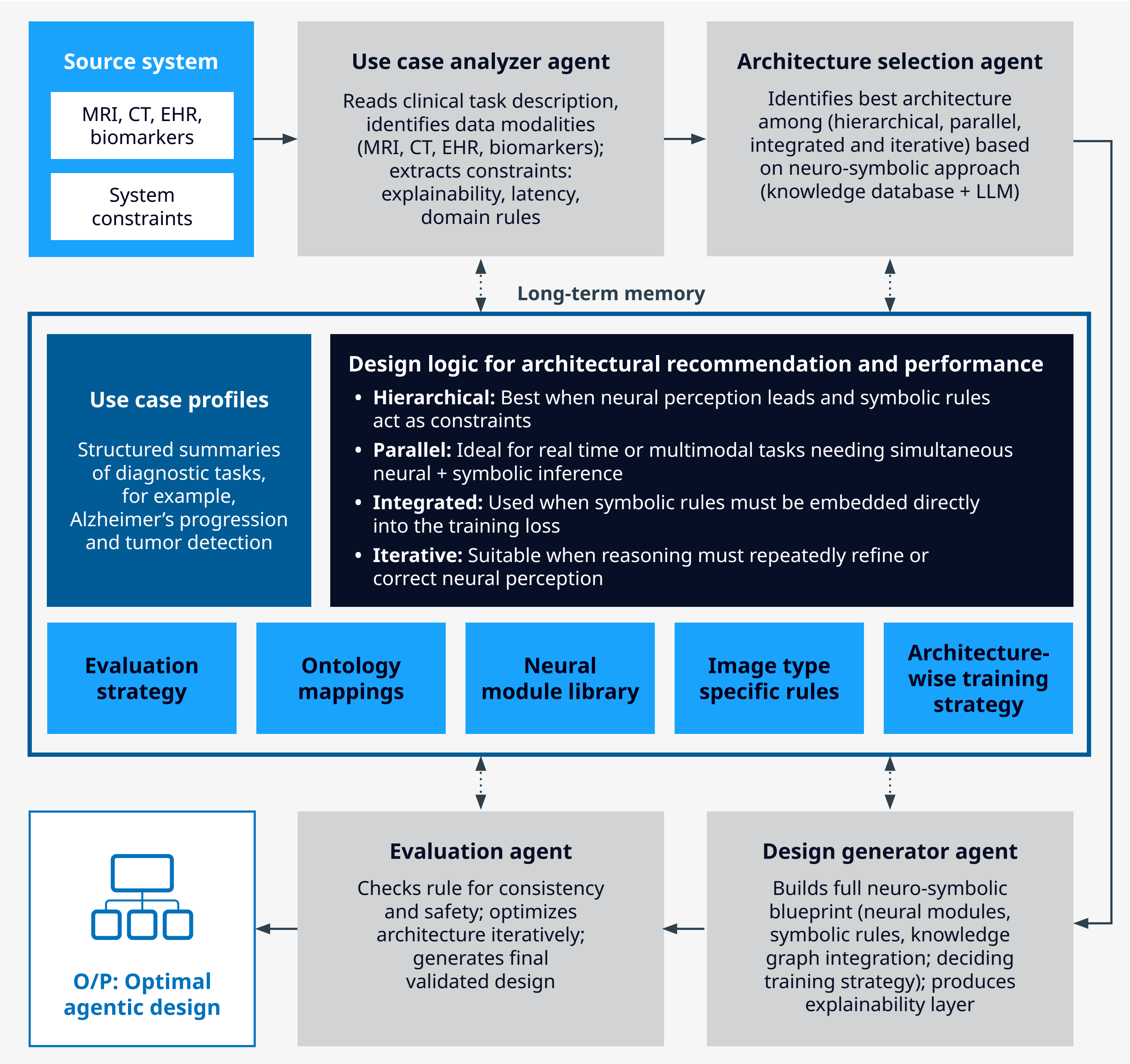


Figure 14: Proposed agentic design for recommending optimal neuro-symbolic architecture

The **Use case analyzer agent** serves as the entry point of the agentic system. Its role is to interpret the clinical problem, understand the data landscape and extract all constraints that influence architectural selection. It performs three core functions:

- **Clinical task interpretation:** The agent reads the diagnostic task description and identifies the nature of the problem (for example, classification, segmentation, progression prediction, multimodal reasoning). It determines whether the task is perception heavy, reasoning heavy or both; this directly influences the choice of neuro symbolic architecture.
- **Data modality identification:** The agent analyzes available data sources such as MRI, CT, PET, ultrasound, EHR text, lab values or biomarkers. It determines the structure, heterogeneity and reliability of each modality, enabling the system to select neural encoders and symbolic components that best match the data characteristics.
- **Constraint extraction:** The agent identifies key requirements and limitations associated with the use case, including the explainability level needed for clinical acceptance; latency or real time processing constraints; domain rules; medical guidelines or ontology dependencies; safety; robustness and regulatory expectations.

These extracted constraints are kept in the use case profiles table/module in the memory that guides the downstream agents in selecting and designing the optimal neuro symbolic architecture.

The **Architecture selection agent** is responsible for determining the most suitable neuro symbolic architecture for a given clinical use case. It acts as the system's decision making core, combining symbolic knowledge (knowledge database for "Design Logic for Architectural Recommendation and Performance") with LLM driven reasoning to evaluate architectural trade offs. The output of this agent is a validated architectural recommendation that guides the downstream Design Generator Agent in constructing the full neuro symbolic system.

The **Design generator agent** is responsible for constructing the complete neuro symbolic system blueprint once the optimal architecture has been selected. It operationalizes the design by assembling neural components, symbolic reasoning layers and training strategies into a coherent, end to end pipeline. Its core functions include:

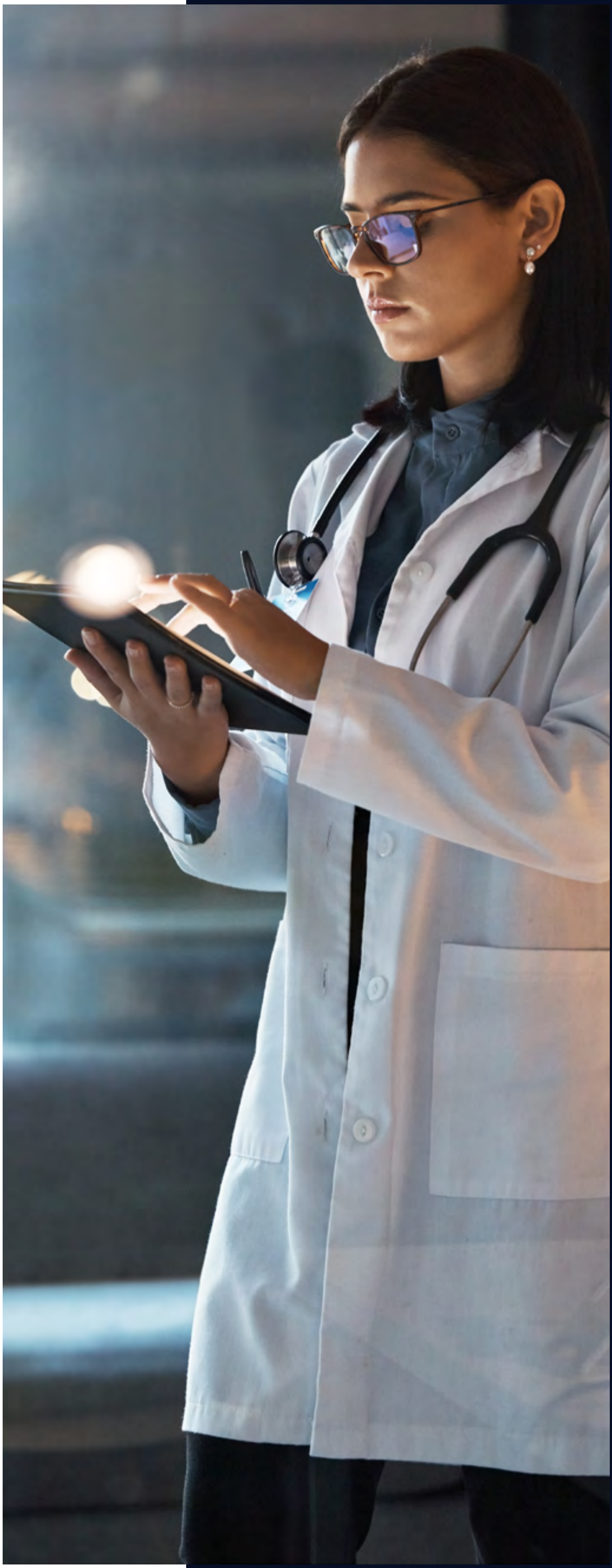
- **Neural module construction:** The agent selects and configures the appropriate neural components (for example, CNNs, ViTs, LLM encoders) based on the data modalities identified in the use case profile.
- **Symbolic rule integration:** The agent incorporates symbolic rules, medical ontologies and domain constraints into the system.
- **Knowledge graph linking:** When the use case requires structured medical knowledge, the agent connects the system to relevant knowledge graphs (for example, UMLS). It defines how neural outputs map to graph entities and how symbolic reasoning queries are executed.
- **Training strategy design:** The agent determines loss functions, semantic penalties, optimization loops aligned with the architecture.
- **Explainability layer generation:** The agent constructs the explainability layer, which produces interpretable outputs such as rule based justifications, reasoning traces, ontology aligned explanations and clinically meaningful summaries. This ensures the final system meets transparency and auditability requirements.

The output of the Design generator agent is a complete, machine generated neuro symbolic blueprint that is ready for evaluation, refinement and eventual deployment in clinical workflows.

The **Evaluation and refinement agent** is the final decision making layer of the agentic system. It ensures that the final neuro symbolic system is accurate, safe, explainable and aligned with clinical standards, completing the agentic design workflow.

The **Long term memory layer** serves as the system's persistent knowledge base, storing reusable information such as use case profiles, symbolic rule templates, ontology mappings, neural module libraries, and architecture specific training strategies. It enables the agentic system to recall prior designs, leverage domain knowledge and maintain consistency across tasks. By preserving performance history and evaluation strategies, it supports continuous self improvement and more accurate architecture recommendations.

The key proposition of this work is the establishment of a higher level agentic system that can dynamically decide which neuro symbolic architecture to select based on the use case. Unlike static frameworks, this system evaluates functional requirements such as accuracy, speed, scalability and explainability, while also considering non functional requirements like latency, robustness and regulatory compliance. By embedding governance rules and enabling human in the loop oversight, the system ensures that architectural recommendations remain clinically safe, auditable and aligned with standards such as FDA and CE. This approach is unique because it transforms architecture selection from a matter of fact choice into an adaptive, intelligent process that continuously learns and improves. In summary, the novelty of this framework lies in its agentic design — a self improving, regulation aware and human guided system that advances neuro symbolic AI from static design patterns to dynamic, trustworthy clinical deployment.



8. Applications in medical imaging

Healthcare will benefit greatly from neuro-symbolic AI. The adoption of NSAI can lead to diagnostic and treatment recommendations that are accurate, explainable and aligned with medical knowledge.[15]

Applications in this domain include.

- **Disease detection and classification:** Researchers have applied Logical Neural Networks (LNNs) for diagnosis prediction systems, with the objective of developing explainable diagnosis models.
- **Explainable diagnostic decision support:** Another important use case is clinical decision support for treatment recommendations. Here, a system might ingest a patient's records (including lab results, symptoms and history) via neural NLP models, then apply symbolic medical guidelines or an ontology (such as SNOMED CT or treatment protocols) to arrive at a diagnosis and advice.
- **Domain ontologies and knowledge bases:** These play a central role in medicine, encompassing symptoms, diagnoses (using ICD codes), treatments and anatomical information. Neuro-symbolic methods naturally make use of this structured knowledge. A neuro-symbolic clinical decision support system (CDSS) might, for instance, use a knowledge graph of drug interactions to flag unsafe medication combinations recommended by a neural model.
- **Medical report generation with reasoning:** Explanations in healthcare often use causal or logical language, for example, "We suspect diabetes because the patient has high fasting glucose, frequent urination and obesity." Where a deep network might only return a probability score without any context, neuro-symbolic models can generate those explanations by tracing their symbolic reasoning steps. Providing clear, step-by-step reasons builds trust with physicians who need to verify AI recommendations and makes it easier to spot mistakes when a reasoning path reveals a misinterpreted symptom.
- **Disease progression analysis:** In disease progression analysis, neuro symbolic methods integrate neural models that extract clinically relevant signals from unstructured text with symbolic medical knowledge that provides structured interpretation and temporal reasoning. This hybrid approach enables accurate and explainable tracking of how a disease evolves over time, allowing clinical events such as progression, remission and metastasis to be consistently mapped to standardized concepts and aligned into coherent longitudinal trajectories.
- **Personalized medicine:** Neuro symbolic analysis strengthens personalized medicine by combining data driven learning with explicit medical reasoning. Through causality aware modeling, systems such as Delphi use causal graphs and discretized physiological variables to explain why a treatment benefits a specific patient rather than merely predicting that it will. Rule guided reasoning ensures that clinical guidelines and safety constraints are encoded symbolically, preventing unsafe or non compliant recommendations. This leads to interpretable decisions in which the system provides structured justifications for each action, improving clinician trust. Finally, dynamic personalization allows neuro symbolic reinforcement learning to adapt treatment strategies in real time as patient states evolve, an approach already demonstrated in complex scenarios such as sepsis management.

[15] He, Y., Wang, Z., Zhang, Y., Dan, T., Chen, T., Wu, G., & Li, A. (2025). NeuroSymAD: A Neuro-Symbolic Framework for Interpretable Alzheimer's Disease Diagnosis. arXiv preprint arXiv:2503.00510.

9. Conclusion

Neuro symbolic AI brings together the strengths of neural networks and symbolic reasoning to create medical AI systems that are both accurate and explainable. By combining pattern recognition with clinical rules and medical knowledge, these systems overcome the limitations of black box deep learning and provide decisions that clinicians can trust. Across imaging, diagnosis and clinical NLP, neuro symbolic methods show clear improvements in transparency, robustness and alignment with real medical reasoning. As healthcare moves toward safer and more accountable AI, neuro symbolic approaches offer a practical path to building diagnostic tools that not only predict outcomes but also clearly explain why — supporting better decisions and improving patient care.

However, it is important to acknowledge the trade offs of neuro symbolic architectures. Integrating symbolic reasoning with neural models can introduce additional computational overhead, sometimes resulting in latency that may affect real time diagnostic workflows. The complexity of maintaining knowledge graphs and logical rules also adds to system design challenges, and scalability may be lower than that of purely neural approaches. Thus, while NSAI offers transparency, robustness and regulatory alignment, these benefits must be balanced against performance and costs to ensure practical clinical deployment.

10. References

Ejalonibu, A. K. (2025). Neuro Symbolic Architectures with Artificial Intelligence for Collaborative Control and Intention Prediction.

Blog on “Neuro-Symbolic AI for Multimodal Reasoning: Foundations, Advances, and Emerging Applications”

He, Y., Wang, Z., Zhang, Y., Dan, T., Chen, T., Wu, G., & Li, A. (2025). NeuroSymAD: A Neuro-Symbolic Framework for Interpretable Alzheimer’s Disease Diagnosis. arXiv preprint arXiv:2503.00510.

Prenosil, G. A., Weitzel, T. K., Bello, S. C., Mingels, C., Manzini, G., Meier, L. P., ... & Afshar-Oromieh, A. (2025). Neuro-symbolic AI for auditable cognitive information extraction from medical reports. Communications Medicine, 5(1), 491.

García-Barragán, Á., Sakor, A., Vidal, M. E., Menasalvas, E., Gonzalez, J. C. S., Provencio, M., & Robles, V. (2025). NSSC: a neuro-symbolic AI system for enhancing accuracy of named entity recognition and linking from oncologic clinical notes. Medical & biological engineering & computing, 63(3), 749-772.

Feldstein, J., Jurčius, M., & Tsamoura, E. (2023, July). Parallel neurosymbolic integration with Concordia. In International Conference on Machine Learning (pp. 9870-9885). PMLR.

Feldstein, J., Dilkas, P., Belle, V., & Tsamoura, E. (2024). Mapping the neuro-symbolic AI landscape by architectures: A handbook on augmenting deep learning through symbolic reasoning. arXiv preprint arXiv:2410.22077.

Rao, B., Eiers, W., & Lipizzi, C. (2025). Neural theorem proving: Generating and structuring proofs for formal verification. arXiv preprint arXiv:2504.17017.

Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., & De Raedt, L. (2018). Deepproblog: Neural probabilistic logic programming. Advances in neural information processing systems, 31.

Sinha, S., Premsri, T., Kamali, D., & Kordjamshidi, P. (2026). Neuro-symbolic frameworks: Conceptual characterization and empirical comparative analysis. Neurosymbolic Artificial Intelligence, 2, 29498732261443183.

Li, Z., Huang, J., & Naik, M. (2023). Scallop: A language for neurosymbolic programming. Proceedings of the ACM on Programming Languages, 7(PLDI), 1463-1487.

Neuro-Symbolic AI with AllegroGraph

Badreddine, S., Garcez, A. D. A., Serafini, L., & Spranger, M. (2022). Logic tensor networks. Artificial Intelligence, 303, 103649.

Riegel, R., Gray, A., Luus, F., Khan, N., Makondo, N., Akhalwaya, I. Y., ... & Srivastava, S. (2020). Logical neural networks. arXiv preprint arXiv:2006.13155.

List of abbreiations

Abbreviation	Meaning
NSAI	Neuro-Symbolic AI
LLMs	Large Language Models