communications medicine

A Nature Portfolio journal

Article

https://doi.org/10.1038/s43856-025-01194-x

Neuro-symbolic AI for auditable cognitive information extraction from medical reports

George A. Prenosil[1,2,4] ✉, Thilo K. Weitzel[3,4], Sandra C. Bello[2], Clemens Mingels[1], Giulia Manzini[1], Lorenz P. Meier[1], Kuang-Yu Shi[1], Axel Rominger[1] & Ali Afshar-Oromieh[1]

Abstract

Background Large language models (LLMs) such as GPT-4 can interpret free text, but unreliable answers, opaque reasoning, and privacy risks limit their use in healthcare. In contrast, rule-based artificial intelligence (AI) provides transparent and reproducible results but struggles with free text. We aimed to combine the strengths of both approaches to test whether such a hybrid system can autonomously and reliably extract clinical data from diagnostic imaging reports.

Methods We developed a neuro-symbolic AI that connects GPT-4 with a rule-based expert system through a semantic integration platform. GPT-4 extracted candidate facts from free-text reports, while the expert system verified them against medical rules, producing traceable, deterministic labels. We evaluated the system on 206 consecutive prostate cancer PET/CT scan reports, requiring extraction of 26 clinical parameters per report, generating 5356 data points, and answering three study questions: study inclusion, recurrent cancer identification, and prostate-specific antigen (PSA) level retrieval. Outputs were compared against physician-derived references, and discrepancies were reviewed by a blinded adjudicator.

Results Here we show that neuro-symbolic AI outperforms GPT-4 alone and matches physicians in structuring and analysing reports. GPT-4 alone achieves F1 scores of 0.63 for study inclusion and 0.95 for recurrence detection, with 96.6% correct PSA values. Physicians reach F1 scores of 1.00 and 0.99, with 98.1% PSA accuracy. The neuro-symbolic AI scores twice 1.00 with 100% PSA accuracy and delivers always an auditable chain of reasoning. It intercepts two intentionally introduced reports with residual identifiers, preventing unintended transfer of sensitive data.

Conclusions Unlike standalone LLMs, neuro-symbolic AI can safely automate data extraction for clinical research and may provide a path toward trustworthy AI in healthcare practice.

Plain language summary

Medical doctors often write reports as free text, which is hard to reuse for research or care. A large language model is software that reads and writes text by imitating large networks of brain cells. This type of artificial intelligence can extract and organize important information from medical reports. But its reasoning is opaque, answers can be wrong, and it raises privacy concerns. Rule-based artificial intelligence is transparent, responds correctly, and is privacy-protecting but struggles with free text. We combined both artificial intelligence types, so each offsets the other's weaknesses. We tested the system on 206 prostate cancer imaging reports, where it extracted information correctly, showed how it reached its answers, and protected sensitive data. Pairing large language models with rule-based systems could make artificial intelligence safer, more trustworthy, and more useful in healthcare.

Radiology and nuclear medicine diagnostic reports are still dictated as free text, and from them structured, reproducible data must be extracted for clinical trials. Before the advent of large language models (LLMs), natural language processing (NLP) struggled with language ambiguity or with unknown semiosis, whereas LLMs such as the Generative Pre-trained Transformer-4 (GPT-4) already match human cognition across many tasks.

However, three fixed constraints block employing LLMs in healthcare: determinism (answers must not shift with prompt phrasing), traceability (reasoning must be auditable), and confidentiality (protected health data must never leak)[1–5]. Achieving all three is still an open challenge and constitutes a major bottleneck for integrating LLMs into clinical workflows and trials. If it were solved, LLMs could classify clinical reports, match patients to

[1]Department of Nuclear Medicine, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland. [2]Zentit GmbH, Muri bei Bern, Switzerland. [3]SCCE - Scientific Consulting, Computing and Engineering GmbH, Kirchlindach, Bern, Switzerland. [4]These authors contributed equally: George A. Prenosil, Thilo K. Weitzel. ✉e-mail: george.prenosil@insel.ch

trials, and mine unstructured research data[3,6]—capabilities that would hasten the detection of pandemics[7], rare side-effects, or malpractice patterns[8]. We therefore wanted to address the pressing problem how to harness LLMs without sacrificing the above-mentioned healthcare requirements.

LLMs are vast neural networks trained on web-scale general corpora[9] and professional domains, including biomedical literature[10]. Belonging to the branch of *stochastic AI*, LLMs excel at digesting unstructured text, coping with ambiguous phrasing[8,11], and—because it outputs probabilities—reasoning under missing data or uncertainty[12]. Its learned representations transfer readily between tasks, so a single model can pivot from trial matching to guideline summarisation without bespoke re-engineering, unlike the narrow, task-specific programs that characterised classical expert systems, i.e. the *symbolic AI*.

Yet the same design brings healthcare-critical drawbacks[13]. Internal synaptic weights are difficult to explain, so the model's logic is inexplicable to auditors[14]. Pattern matching without formal deduction limits multi-step reasoning[15] and leaves the system blind to out-of-distribution inputs or to its own nescience[16]. Outputs are stochastic and prompt-sensitive, where minor re-phrasings of a semantically identical prompt can lead to divergent answers[17]. This may be a feature for creative writing but a liability when identical clinical facts must yield identical conclusions[18]. Finally, the distributed services of the most capable LLMs, such as GPT-4, raise confidentiality and alignment concerns[5,19].

Compared to neural AI, the older symbolic AI stores knowledge as human-readable symbols and rules in an ontology, and uses formal logic to yield deterministic and reproducible outcomes[20]. This makes inference chains auditable and allows the program to declare 'unknown', ask for some specific data, or trigger other meaningful fall-back, when a rule cannot fire[21]. Early medical expert systems like MYCIN proved this approach in the 1980s[22]. However, symbolic AI is labour-intensive to scale and struggles with unstructured or uncertain inputs, and rapidly evolving medical domains[23,24].

We therefore propose a unified semantic-neuro-symbolic NLP pipeline. The rationale for this approach is that each component offsets the weaknesses of the others: In our design, GPT-4 harvests clinical facts from free text[25], while a locally hosted expert system (Plato-3) verifies these extractions against medical rules, generating deterministic, trustworthy labels[21,26]. Finally, conventional software provides the practical access needed for real-world clinical workflows. In combination, these elements create a system capable of transforming unstructured diagnostic reports into structured, auditable, and privacy-preserving data suitable for research and patient care.

Recent gains in hardware throughput and programming techniques[23] make real-time integration of large neural models and symbolic reasoning systems finally practicable[27]. Interoperability between different AI systems as well as software required for a meaningful healthcare workflow is still hard: Symbolic AI, neural AI, and conventional software all reason at different abstraction levels and data formats[28,29]. To bridge these different representations of a same problem we built RUDS (Rule-based Unification of Digital technology using Semantics). This platform provides 'loose coupling'[30] but high cohesion[31] through semantic message passing between diverse components—also known as semantic unification[32]. RUDS implements multiple programming paradigms, which allows diverse components to exchange and interpret information despite differences in data structure, programming style, or level of abstraction[33], while the embedded expert system keeps track of context and meta information at all time. By elevating all connected components to a shared semantic representation[32], RUDS unites neural cognition with symbolic reasoning in a full semantic-neuro-symbolic AI stack[26], realizing the cognitive computing paradigm[34]. The result pairs LLM–NLP with an auditable inference chain from the expert system[35,36], enabling back-tracing from each final label to the originating LLM tokens or human prompt[37]. This capability is crucial in healthcare, where understanding the 'why' behind AI decisions is essential[38,39].

Our goal was an *exploratory proof-of-concept*, demonstrating seamless cooperation between an LLM and a symbolic expert system in autonomously compiling structured clinical data from free-text diagnostic reports. Specifically, our study makes four practical contributions: First, we integrate GPT-4 with the expert system Plato-3 so that extracted facts are validated by medical rules. Second, each AI-generated label includes the complete symbolic reasoning chain and the supporting GPT-4 evidence, providing explainability-by-design in natural language. Third, we show that the system does not require retraining of the language model; domain knowledge is provided entirely through the rule base. Fourth, the architecture is implemented on the semantic-unification platform RUDS, enabling interoperability with conventional software and realizing the vision of a practical neuro-symbolic clinical AI[3,26,40,41]. In addition, this work also explains the paradigm-unifying architecture of RUDS in detail and why it is needed here. Although our evaluation was modelled after a prior PET/CT clinical study[42], it was not designed to produce new clinical findings. Instead, we show in this proof-of-concept study how the combined neuro-symbolic AI accurately extracts and structures 26 clinical parameters from 206 original, unedited [68Ga]Ga-PSMA-11 reports for recurrent prostate cancer (rPC). The system matches physician performance, outperforms GPT-4 alone, produces deterministic results without hallucinations, and prevents privacy breaches by controlling all data transfer. Taken together, this work demonstrates a practical implementation of the autonomous, context-aware AI originally envisioned by the Japanese fifth-generation computing initiative[43].

Methods

Patient data

The study retrospectively analysed 206 diagnostic reports from 206 consecutive patients who had undergone [^{68}Ga]Ga-PSMA-11 PET/CT scans in eight months between January and August 2018 at the Department of Nuclear Medicine, Inselspital Bern, Switzerland, adhering to Swiss ethical guidelines[44]. The Cantonal Ethics Committee Bern (Kantonale Ethikkommission Bern, Murtenstrasse 31 in 3010 Bern, Switzerland) approved retrospective usage of the patient reports (KEK-Nr. 2018–00299). All patients published in this manuscript signed a written informed consent form for the purpose of anonymized evaluation and publication of their data. No additional approval beyond that was obtained.

We chose [^{68}Ga]Ga-PSMA-11 PET/CT reports because they paired a highly variable free-text narrative—including the patient history—with a compact, guideline-defined decision scheme, while the cohort size remained amenable to iterative human cross-checking. This combination creates a tractable but non-trivial testbed: if the neuro-symbolic pipeline can deliver deterministic answers here, it is well-poised to scale to larger clinical trials that share the same 'unstructured-text + rule set' pattern.' Because the retrospective evaluation of PET/CT reports followed the design of a study published by Afshar-Oromieh et al.[42], we were able to demonstrate real-world applicability, and the experience gained form this study also qualified the authors to develop the expert system's ontology.

The diagnostic reports, originally written and checked by three nuclear medicine physicians unrelated to this study, were formatted into PDF files according to our institutional standards, codified and anonymized using the batch processing software PDF Replacer v.1.8.7.0 (pdfreplacer.com), and split evenly into a development set and a validation set. All reports were again checked manually for correct anonymization. We fabricated two wrongly anonymized sets using an author's name and birthdate for testing the expert system's ability to recognize un-anonymized reports before sending information to the LLM. The reports included 160 patients with rPC, 11 patients having undergone primary tumour staging (PTS), and 29 patients showing no cancer pathology. Two nuclear medicine physicians (C.M., A.A.O.) consensually extracted 26 study-relevant parameters (Table 1) from the 206 reports, providing a physician-generated reference with 5356 data points, e.g. labels. In case a label could not be elicited, both AI systems and the physicians were instructed to record a 'N/A' (not applicable). Inclusion and exclusion criteria were as previously published[42], meaning that rPC, PTS, and non-pathological reports needed to be distinguished. For the study, the 206 reports were split evenly into a development set and a validation set.

Table 1 | 26 study parameters the AI-system was tasked to extract from diagnostic PET/CT reports

Study parameter detection								
Study parameter			GPT-4-only (LLM)		Neuro-symbolic AI		Physician-generated reference	
#	*Label*	*Type*	*Trues*	*Success rate (%)*	*Trues*	*Success rate (%)e*	*Trues*	*Success rate (%)*
1	Anonymization[a]	Yes/No	n/a	n/a	206	100	206	100
2	Patient name[b]	String	206	100	206	100	206	100
3	Patient age[b]	Integer	206	100	206	100	206	100
4	Exam date[b]	Date	206	100	206	100	206	100
5	PSA value[b]	Number	202	98.1	206	100	199	96.6
6	PSA date[b]	Date	200	97.1	206	100	196	95.1
7	Gleason score[b]	String	206	100	206	100	199	96.6
8	Pre-therapy[a,b]	Yes/No	204	99.0	206	100	203	98.5
9	Primary tumour staging[a,b]	Yes/No	200	97.09	206	100	205	99.51
10	Pathological report[a,b]	Yes/No	189	91.75	206	100	204	99.03
	Total success[b]		1819	98.1 ± 2.7	2060	100.0 ± 0	2030	98.5 ± 1.9
11	TNM	String	189	91.7	n/a	n/a	190	92.2
12	Ongoing ADT	Yes/No	204	99.0	n/a	n/a	203	98.5
13	Radical prostatectomy	Yes/No	203	98.5	n/a	n/a	197	95.6
14	Primary tumour recurrence	Yes/No	205	99.5	n/a	n/a	204	99.0
15	Primary tumour quantity	Integer	199	96.6	n/a	n/a	202	98.1
16	LNM occurrence	Yes/No	204	99.0	n/a	n/a	205	99.5
17	LNM quantity	Integer	147	71.4	n/a	n/a	201	97.5
18	LNM locus	String	203	98.5	n/a	n/a	201	97.5
19	Bone metastases occurrence	Yes/No	203	98.5	n/a	n/a	206	100
20	Bone metastases quantity	Integer	187	90.8	n/a	n/a	203	98.5
21	Bone metastases locus	String	205	99.5	n/a	n/a	206	100
22	Organ metastases occurrence	Yes/No	199	96.6	n/a	n/a	205	99.5
23	Organ metastases quantity	Integer	200	97.1	n/a	n/a	205	99.5
24	Organ metastases locus	String	200	97.1	n/a	n/a	204	99.0
25	Secondary tumour occurrence	Yes/No	166	80.6	n/a	n/a	206	100
26	Secondary tumour locus	String	177	85.9	n/a	n/a	206	100
	Total success overall		4910	95.3 ± 6.8	2060	100.0 ± 0	5274	98.4 ± 1.9

[a]Meta-parameters determined through rule based reasoning from other parameters.
[b]Parameters covered by all AI entities examined. All detection rates refer to the retrospectively, manually, retrospectively corrected, "reviewed reference". *ADT* androgen deprivation therapy, *AI* artificial intelligence, *GPT-4* generative pre-trained transformer 4, *LLM* large language model, *LNM* lymph node metastases, *PSA* prostate-specific antigen, *TNM* Classification of malignant tumours with Tumour size (T), lymph Nodes (N), and distant Metastases (M); n/a: Parameter was not examined by that entity. Averages are shown with ± standard deviation.

Paradigm-integrating platform - theory and implementation

We combined an LLM, an expert system and conventional software into a semantic neuro-symbolic AI system running on our study software. Table 2 summarises how each system component compensates for the other's limitations in this setup.

The study software was developed and operated on RUDS (Zentit GmbH, Muri bei Bern, Switzerland; ruds.ch), written in Java™8 (Oracle, Austin, TX, USA) on Netbeans™IDE 8.2 (Apache Software Foundation, Wilmington, DE, USA) on a Dell Precision 5470 laptop (Dell Inc., Round Rock, TX, USA) with Microsoft Windows™10. GPT-4 integration required internet access and an API account from OpenAI Global, LLC (San Francisco, CA, USA). The temperature setting of GPT-4 was left on its default value of one when running the analysis in May 2024.

RUDS's software architecture enables seamless interaction between diverse computational paradigms of software, thereby removing dependencies in technology, instruction semantics, data, and process order. It integrates multiple programming paradigms representing various levels of abstraction (Fig. 1a) enabling the software to address the entirety of a complex problem[33]. The multiple levels of abstraction[28] reflect real-life systems, processes or workflows (Fig. 1b) as needed. As such, process abstraction is the act of converting a process description into a more abstract form, resulting in a decrease in model components, interactions, and behavioural complexity. This allows for a higher-level representation that captures core ideas or functionality[29]. Thus, the choice of programming paradigm influences how software processes data and determines, with the addressable abstraction level, the addressable complexity and context of a problem[45]. Abstraction programming bridges complexity gaps between system components[45], by '*gluing meanings of parts of a discourse into a coherent whole*'[32] and retains context across abstraction levels[46]. Merely applying computational methods without considering contextual meaning[11] of data will lead to meaningless results[47], which can endanger patient safety and proper study outcomes.

LLMs usually require specific APIs to couple with non-semantic information technology. Our top semantic abstraction level with its bidirectional semantic information flow enables high-cohesion loose coupling[30,31] of semantic and non-semantic information technologies and serves at the same time to inform the human user about the system's processes. It also enables 'interface reversal,' where LLMs and expert systems can either actively engage with and utilize any connected application modules, or the application modules themselves can access AI capabilities. For instance, conventional software can leverage AI for advanced tasks, such as data interpretation or

Table 2 | Strengths of one system component balance weaknesses of other system components in a semantic-neuro-symbolic AI setup

COMPARISON OF NEURAL AI, SYMBOLIC AI AND CONVENTIONAL SOFTWARE		
System component	**Strengths**	**Weaknesses**
1. Neural, stochastic AI: GPT-4	+ Accepts unstructured data	- Logical reasoning
	+ Accepts ambiguity	- Factuality
	+ Generalizable	- Transparency
	+ Huge knowledge	- Privacy & alignment
	+ Good scalability	- Limited interoperability
		- Limited real-world access
2. Rule-based symbolic AI: Plato-3	+ Logical reasoning	- Struggles with unstructured data
	+ Factuality	- Struggles with ambiguity
	+ Transparency	- Works domain specific
	+ Privacy & alignment	- Limited knowledge
	+ Good interoperability	- Limited scalability
	+ Semantic data representation	- Limited real-world access
3. Conventional software: File loader, PDF-reader, Report generator	+ Provides real-world access	- "Dumb", non-semantic
	+ Availability	- Limited Interoperability
	+ Human centric	- Works domain specific
	+ Privacy & alignment	- Uses legacy data formats

AI artificial intelligence, *GPT-4* generative pre-trained transformer 4, *PDF* portable document format.

decision support, while AI can call semantically upon these tools for operations like data formatting or numerical processing.

To archive this, all application modules of the study software communicate semantically through the paradigm-integrating platform. For this, the platform uses a universal messaging protocol, recursive universal objects, and a central 'blackboard'[48] acting as a shared information space and common ground[49]. The seamless semantic communication across paradigms (Fig. 1a) enables e.g. numerical procedures working on a low level of abstraction to ask for support from higher levels of abstraction, e.g. the expert system, for making decisions about details of their own use in a self-referencing procedure[50]. Algorithms at the procedural abstraction level can therefore adapt at runtime to a given context or to the intent of a computation that can only be inferred at a higher level of abstraction.

Vice versa, processes on a higher level of abstraction can easily access functionality of their choice from lower levels to obtain information they need or to process data as needed for reaching a given goal. Consistently, the platform uses the same design patterns for cross-paradigm access to a graphical user interface or even to APIs for third-party applications such as GPT-4. Information coming from or being send to other digital applications and databases is always first handled at its native abstraction level and then made accessible to applications working at lower or higher levels.

The application modules handle tasks such as text extraction, data preprocessing, and report generation (Supplementary Table S1). All modules exchange information along with meta-information[51], such as AI reasoning and data origin. Unlike in most conventional software, information and associated meta-information persists and remains accessible to all subsystems at all abstraction levels, even when the subsystems themselves lack semantic capabilities. Finally, the 'agent' software design pattern[52] and the data, context, and interaction paradigm[53] bestow agency onto parts of the system. Supplementary Table S1 explains the various concepts of the software.

Expert system implementation

The NLP expert system Plato-3 in Prolog[21] is an integral part of RUDS. Plato-3 communicates semantically with GPT-4 and the other required application modules, and its rule base realizes the ontology software paradigm[54]. Plato-3 speaks a distinct German subset of natural language for oncologic PET reports, using the definite clause grammar formalism[55]. The recursively enumerable language spans the full Chomsky hierarchy of formal grammar classes[56]. Consequently, Plato-3 accepts unknown words from GPT-4 but provides a defined language set to GPT-4 for unambiguous communication between the two different AI concepts. This semantic unification[32] ensures that the two AI systems, despite operating with different internal representations and reasoning methods, share a common semantic framework for consistent and unambiguous communication, thereby realizing the neuro-symbolic paradigm (Fig. 1a).

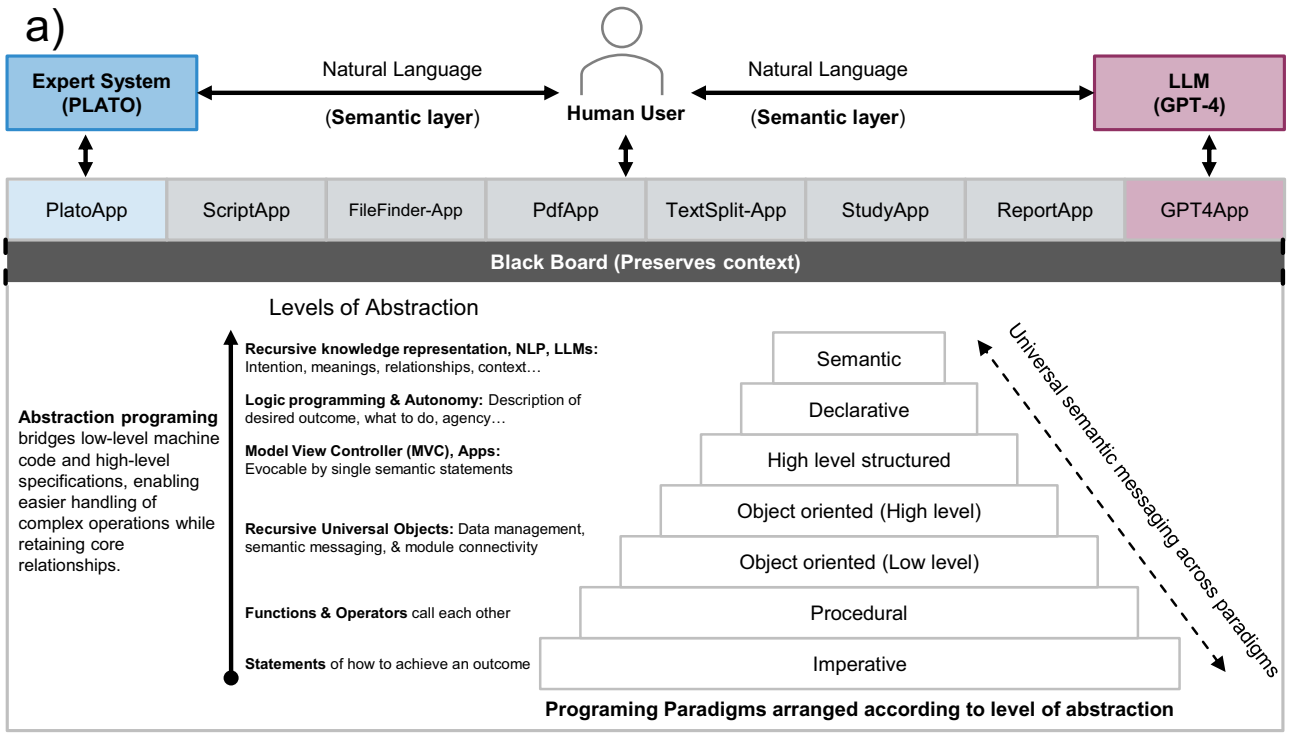
Plato-3's ontology contains facts along with their logical relations, rules, and problem-solving methods[20]. In logic, a fact is a statement that is unequivocally true within a given domain, such as 'Patient X's report mentioned a relapse.' A rule is a conditional statement that connects facts and enables reasoning or inference, such as 'If a relapse is mentioned, then the PET/CT report is classified as 'no PTS'.' These constructs allow Plato-3 to use recursive first-order predicate logic for reasoning and deriving new facts and knowledge from already existing facts[47].

Here, the ontology encompasses clinical guidelines for anonymization, study inclusion (Fig. 2), identifying pathological reports, and determining correct prostate-specific antigen (PSA) values. The ontology can integrate new facts and rules from various modules, expressed in natural or symbolic language, enabling self-modification[50].

Study aims and workflow

We tasked the system with extracting 26 study parameters (Table 1) from the manually pre-anonymized and codified 206 PET-reports and answer three main study questions, relevant for patient inclusion and aims of the reference study[42]. The first study question concerned study inclusion: Does the properly anonymized report describe rPC after radical prostatectomy or PTS (Parameter 9)? The second study question concerned identifying reports mentioning pathology: Was a pathology found by the PSMA PET/CT, was it rPC, and how many tumour locations were found (Parameter 10)? The third study question concerned PSA-levels mentioned in the reports: What was the PSA-level measured at the time closest before the PET/CT scan, and what was the time interval between the PSA and the PET scan (Parameters 5 and 6)?

The answers to these three main study questions could not be simply parsed from the reports by the AI but required inference from other study



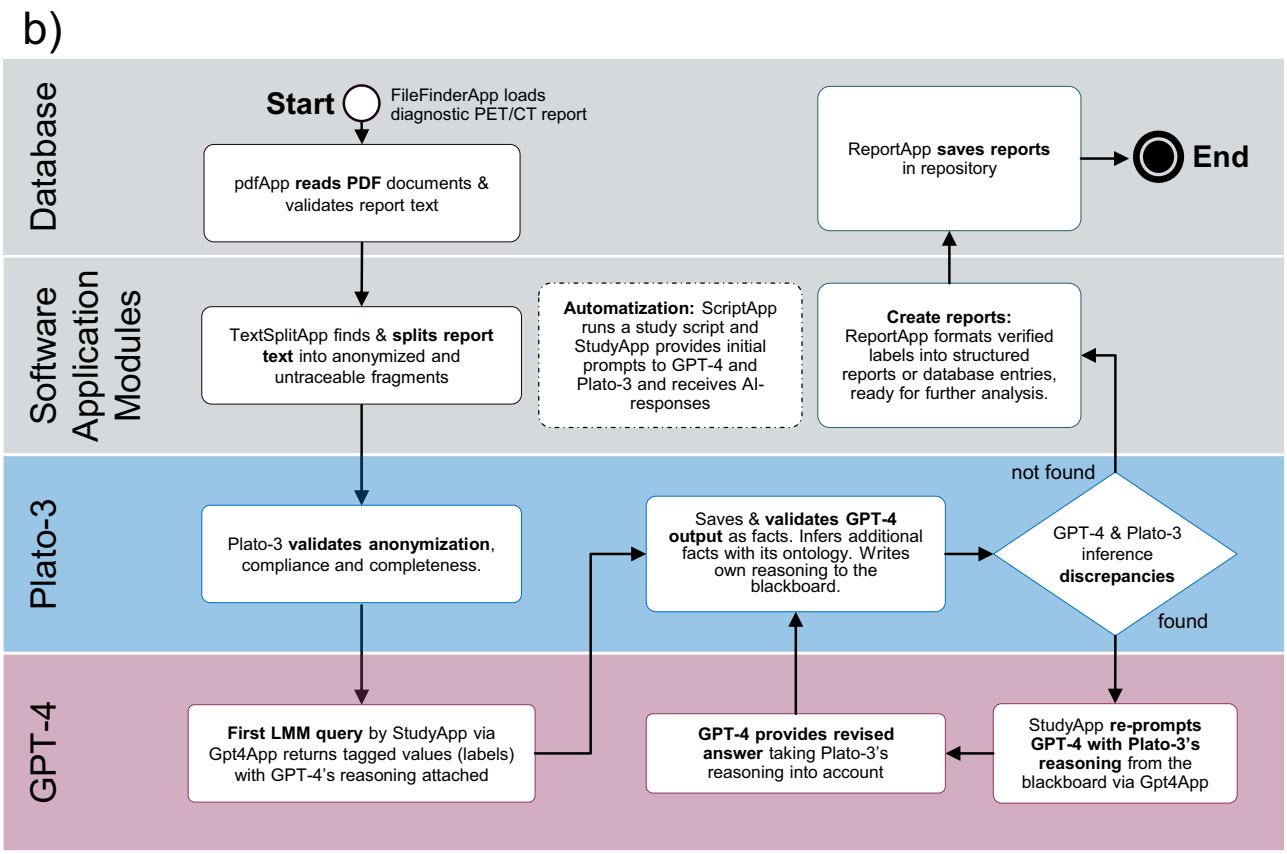


Fig. 1 | Software architecture and workflow. a Paradigm integrating software architecture showing the programming paradigms (stepped pyramid) used at the respective abstraction level and the software applications (Apps, coloured boxes) used in the workflow. Paradigm integration enables interfacing different AI and software applications for specific workflows into a single system. The universal semantic messaging (Dashed arrow) that ensures information flow across abstraction levels and the blackboard (Black bar) enable generic information exchange between connected components. **b** Workflow diagram for analysing a single diagnostic PET/CT report. White boxes depict stations in the workflow, where the applications (grey bars) shown in **a** performed their specific tasks. The black arrows show the flow of data and information between the applications, Plato-3 (blue bar), and GPT-4 (pink bar). The workflow looped for each individual report once. GPT-4 generative pre-trained transformer 4, PDF portable document format, NLP natural language processing.

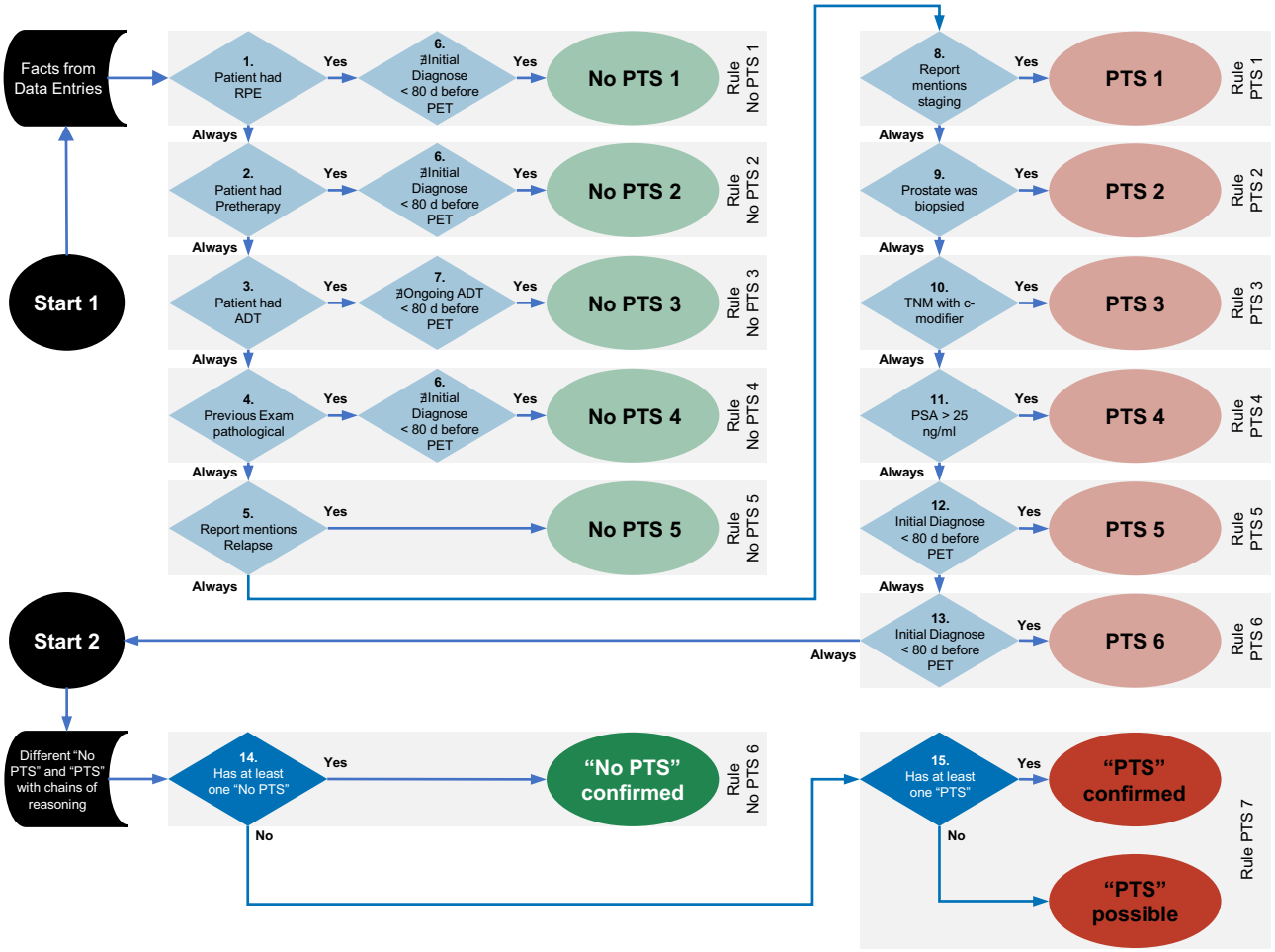


Fig. 2 | Simplified excerpt of the Plato-3 ontology used to detect PTS reports. **Start 1** yields an initial set of meta-facts, which are re-checked after **Start 2**. Rhomboids depict fact checking within a rule. Light-green/red nodes mark intermediate reasoning states that create additional facts; dark-green/red nodes mark final verdicts. Fall-back rules are omitted for clarity but would result in *"PTS status could not be established"* with an accordant reasoning (such as *"data missing"*). *Example – Rule "No PTS 1": If the patient has undergone radical prostatectomy and no rPC was diagnosed initially in the 80 days before the PET examination, primary-tumor staging is ruled out. In contrast, Rule "No PTS 5" reads: If rPC was diagnosed initially in the 80 days before the PET examination, PTS is still possible.* The final confirmation would then be done here with rule PTS7. ADT androgen deprivation therapy, PET positron emission tomography, PSA prostate-specific antigen, rPC recurrent prostate cancer, RPE radical prostatectomy; TNM = Classification of malignant tumours with tumour size (T), lymph nodes (N), and distant metastases (M), c-modifier indicates that stage was tumour stage was determined before treatment.

parameters. To this end, Plato-3 inferred study parameters 1 to 10 from logical relations between already detected study parameters, while extracting parameters 11 to 26 primarily involved parsing and structuring text by GPT-4 and did not require expert system inference (Table 1). This proceeding reflects the division of labour between the two different AI types.

Before sending data to GPT-4, the workflow (Fig. 1b) included the TextSplitApp application module segmenting every report into a clinical history with clinical question, the clinical findings, and the physicians' conclusions and Plato-3 verifying anonymization. This ensured that the resulting fragments passed to GPT-4 contained no trackable or sensitive information. Then the software used pre-engineered prompts to guide GPT-4 in data extraction. We refined these pre-engineered prompts as well as Plato's ontology (Fig. 2) through iterative testing on the development dataset. Most prompts asked GPT-4 to provide a reasoning statement at the end of its answer. Plato-3 saved GPT-4's answers together with its reasoning statements in its database, converted them into semantic facts, and checked these facts against its rules to create new facts. When finding controversial facts, Plato-3 re-consulted GPT-4 with its own reasoning. The primal reasoning of GPT-4 served hereby as a starting point for Plato's chain of thoughts. Even though, Plato-3 could not verify GPT-4's primal reasoning, it could rule out controversial facts and therefore accept only reasoning from plausible fact statements as its primal reasoning. All in all, the expert system takes on the role of a quality control reviewer, verifying inputs to and outputs from the LLMs, checking data against clinical rules and study guidelines, and flagging any inconsistencies or breach of rule with clear reasoning[41]. Supplementary Table S3 provides a detailed description of how the software's core components direct the workflow.

Statistics and reproducibility

First, we compared output from the neuro-symbolic AI, e.g. GPT-4 combined with Plato-3, to the physician-generated reference established by two physicians and one study nurse. Discrepancies between outputs were manually reassessed by authors A.A.O. and G.A.P. in the respective PET reports, using the AI's reasoning chain to identify the underlying cause for conflict. This human-driven reassessment generated a new reviewed reference, which served as the new ground-truth for comparing outputs from GPT-4 alone (GPT-4-only), neuro-symbolic AI, and the original physician-generated reference.

To avoid circularity in preparing the ground truth, the adjudication pipeline followed three steps. Dual encoding of every report: an original label was supplied by the reporting physician, while two AI labels were produced independently by the neural (GPT-4-only) and neuro-symbolic pipelines.

Blinded adjudication (A.A.O.): all discordant items were re-read and re-labelled by a nuclear medicine physician (A.A.O.) who was blinded to which label came from the AI. Logic concordance check (G.A.P.): the adjudicator's decision and the AI's explicit rule trace were then reviewed by a second author (G.A.P.). A correction was accepted only if the adjudicator's reasoning and the AI's rule-based trace were fully congruent; otherwise, the original human label was retained.

Each of the 206 PET/CT reports represented one independent sample. For parameter extraction, every report contributed one replicate per parameter, yielding 5356 data points in total. Replicates are therefore defined as individual parameter extractions from independent reports. Each report was analysed once by GPT-4, once by physicians, and once by the combined neuro-symbolic AI. No technical replicates or repeated runs were used.

Sensitivity, specificity, and predictive performance (F-score) were evaluated in discriminating pathological cases and identifying PTS patients for the three comparisons. McNemar's test with exact binomial testing to calculate two-sided p-values, with $p \leq 0.05$ considered statistically significant. Agreement with the reviewed reference in identifying the remaining study parameters was reported as a percentage, with the Pearson's chi-squared multinomial test applied on parameters 2 to 10, which were shared between GPT-4-only, neuro-symbolic AI and the physician-generated reference (Table 1). Bonferroni correction was used to correct for multiple measurements.

Results

Reviewed reference

Plato-3 blocked the two intentionally wrong anonymized data sets; one with the author's name and one with the birthdate in plain text. Furthermore, it flagged 17 reports missing disclaimer text. The latter indicated possibly missing written consent, which was retrieved in all cases after manual rechecking. The following manual inspection of discrepancies between the physician-generated reference and AI results revealed a total of 82 human errors (Table 1). Furthermore, the neuro-symbolic AI's reasoning in natural language (Example given in Fig. 3) was consistently accurate, while GPT-4-only was not. Therefore, all 32 changes in study parameters 1 to 10 suggested by the combined neuro-symbolic AI were accepted into the reviewed reference, while only 52 out of 207 changes in parameters 11 to 26 suggested by GPT-4-only were accepted here. Since no performance differences were observed between the development and validation sets, they were combined for the results. GPT-4 always agreed with Plato-3's chain of reasoning, when re-prompted (*c.f.* Step c in Fig. 3).

Study inclusion

GPT-4-only correctly excluded two PTS patients from the development set and three from the validation set but missed three PTS patients in each set. There were no false positive PTS categorizations by GPT-4, and it achieved a sensitivity of 0.45, specificity of unity, and F-score of 0.63. Under Plato-3's supervision, the neuro-symbolic AI correctly excluded all 11 PTS patients from the cohort, bringing sensitivity, specificity, and F-score to unity. Manual inspection revealed that the neuro-symbolic AI correctly identified one PTS patient mislabelled as rPC by physicians, giving a sensitivity of 0.91, specificity of unity, and F-score of 1 (Table 3).

Table 4 lists each PTS patient that GPT-4 missed. Observed failure patterns were (i) Trigger-word bias (ii) Missing temporal reasoning (iii) Hallucinated clinical context when phrasing was ambiguous. Because Plato-3 re-evaluates every structured fact against explicit guidelines and stored facts, all six misclassifications were corrected, so the combined neuro-symbolic system produced the right PTS label in every case.

Pathological reports

Without Plato-3, GPT-4-only correctly classified 176 PET/CT reports as pathological and 13 as healthy but misclassified 16 cases as pathological and 1 case as healthy, resulting in a sensitivity of 0.99, specificity of 0.45, and F-score of 0.95. In all 16 false positive cases, GPT-4 answered 'no' to specific questions about primary tumours or metastases but speculated a pathology due to clinical history. The neuro-symbolic AI made no mistakes and achieved a sensitivity, specificity, and F-score of unity. In the previously 16 false positive cases, GPT-4 reverted its opinion after receiving Plato-3's accordant chain of reasoning. Furthermore, the neuro-symbolic AI identified one false negative and one false positive pathology classification in the physician-generated reference, resulting in a sensitivity of 0.99, specificity of 0.97, and F-score of 0.99 (Table 3).

PSA-level and other study parameters

GPT-4-only made four mistakes in the development set and zero mistakes in the validation set when identifying the latest PSA-level, resulting in an overall agreement with the reviewed reference of 98.1%. With Plato-3 re-prompting GPT-4, the neuro-symbolic AI improved agreement to 100%. Mistakes in identifying the correct PSA-level made by GPT-4 were exclusively due to erroneously formatted or written PET reports. Examples include missing measurement units, non-attributable dates, or incorrect designations. These ambiguities affected also the human readers, as the neuro-symbolic AI uncovered seven PSA-level mistakes made by the physicians, giving a human agreement rate of 96.6%. The overall agreement for correctly detecting study parameters compared to the review standard was 94.7 ± 7.1% for GPT-4-only and 98.4 ± 1.9% for the physician-generated reference. The agreement for correctly detecting study parameters that were covered by the ontology (Parameter 2 to 9) was 98.1 ± 2.7% for GPT-4-only, 98.4 ± 1.8% for the physician-generated reference, and 100 ± 0% for the neuro-symbolic AI (Table 1). No significant differences could be observed between the data sets.

Discussion

Our main result is the proof of concept for our approach to realize neuro-symbolic AI. At autonomously structuring and analysing medical reports the neuro-symbolic AI outperformed the unaided LLM (GPT-4) and matched or outperformed trained physicians. GPT-4 alone performed similarly to previously published results that showed over 90% success in text mining[6]. However, when extracting study parameters controlled by the expert, the combination of GPT-4 and the expert system reached near-perfect accuracy. This was especially apparent for the three main study questions, where inference was required. The expert system's oversight also ensured confidentiality, reducing privacy risks posed by GPT-4's distributed nature.

Our results explore how a semantic-neuro-symbolic AI can discern and communicate complex factual issues, such as identifying distinct patient groups or necessary facts for the replication of the reference study[42]. The correct identification of PTS constituted the most demanding task, requiring the AI to construct multiple fact layers. Checking facts against a dedicated ontology and using a specific subset of natural language removes ambiguity from the workflow and enables GPT-4 to structure PET reports into validated data without specialized training for the task. The human operator is always able to retrace step-by-step the neuro-symbolic AI reasoning for every label, as well as to understand and correct his own mistakes. Figure 3 demonstrates this traceability and shows together with Table 4 how the expert system complemented the LLM in the decision-making process.

A key strength is that we retained every PET/CT report exactly as dictated by three different staff physicians, each using their own phrasing, abbreviations and section layout; this heterogeneity shows that the neuro-symbolic pipeline handles natural language that extends well beyond a single author-specific template. Additionally, using the publicly accessible GPT-4 and not a specifically trained foundation model makes our findings generalizable to other LLMs.

Artificial intelligence will be introduced in healthcare, and decision traceability and fact checking becomes crucial due to the high importance of accountability. Especially, the correct identification of wrongly anonymized data and missing disclaimers highlights the need for prudence when using stochastic 'black box' technologies like LLMs. This caution arises from the potential risks to patient wellbeing and privacy due to opaque decision-making processes and unclear data handling by unsupervised autonomous software[19].

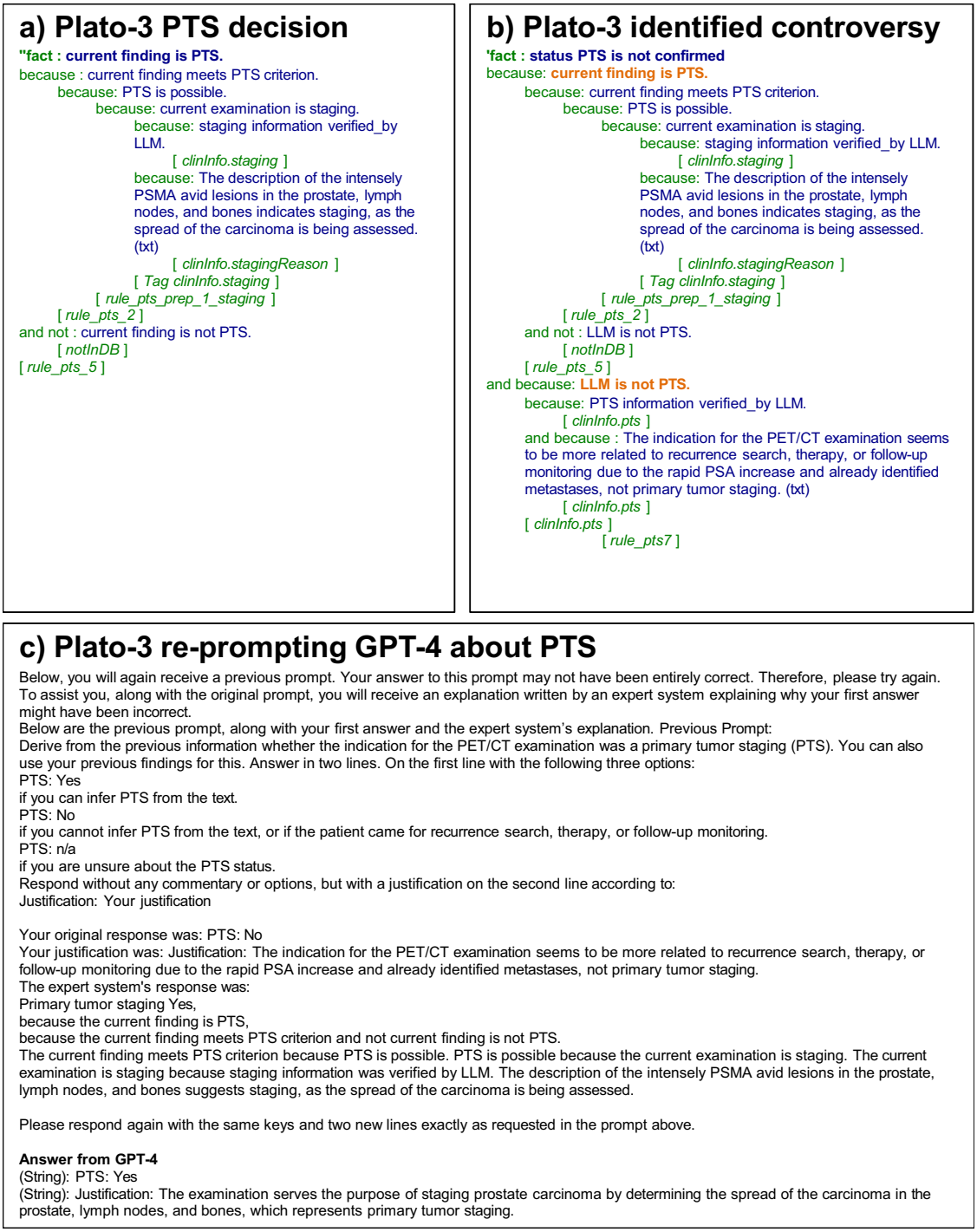


Fig. 3 | Traceable reasoning process in four steps shown on the example of detecting a PTS-report (Patient 15) as presented to the user. a First step with trace of Plato-3 decision. **b** Second step with trace where Plato-3 identified a controversy between its decision and the decision of GPT-4. **c** Third step showing the trace with which Plato-3 re-prompted GPT-4 with the identified controversy and its decision. The answer from GPT-4 is the fourth step. Green: Plato-3's predicates, including the information sources or applied rules in squared parentheses; [notInDB]: Plato-3 could not find the fact in his database. txt: The text contains language elements unknown to Plato-3. Blue: Plato-3's reasoning; The originally German text was translated to English, replicating Plato-3's distinct set of natural language. LLM large language model, PET/CT positron emission tomography combined with computed tomography, PSMA prostate-specific membrane antigen, PTS primary tumour staging, n/a not available.

Most approaches to combat a LLM's stochastic output focus on either fine-tuning the model with specific medical data[57], giving access to external databases, or embedding symbolic knowledge directly into neural networks[27,58], which limits these systems to specific, narrowly-defined tasks. In contrast, our validation study uses the RUDS platform that enables interaction between different programming paradigms and thus the generic interfacing of different AI concepts into a single neuro-symbolic system. Instead of tightly coupling symbolic AI within the neural network, RUDS allows symbolic and neural engines to prompt each other dynamically and call upon additional AI models or traditional software as needed. This broadens the scope of problems such a combined AI can address, moving

Table 3 | Comparison of GPT-4-only, Neuro-symbolic AI, and Physician-generated reference for discerning primary tumour staging (PTS) reports and pathological reports mentioning recurrent prostate cancer (rPC)

Confusion matrices for discerning PTS reports and pathological reports									
	Study question	GPT-4-only			Neuro-symbolic AI			Physician-generated reference		
Reviewed standard	PTS	206 T	5 PP	201 PN	206 T	11 PP	195 PN	206 T	10 PP	196 PN
		11 P	5 TP	6 FN	11 P	11 TP	0 FN	11 P	10 TP	1 FN
		195 N	0 FP	195 TN	195 N	0 FP	195 TN	195 N	0 FP	195 TN
	Pathological report (rPC)	206 T	192PP	14 PN	206 T	177 PP	29 PN	206 T	177 PP	29 PN
		177 P	176 TP*	1 FN*	177 P	177 TP	0 FN	177 P	176 TP	1 FN
		29 N	16 FP*	13 TN*	29 N	0 FP	29 TN	29 N	1 FP	28 TN

FN false negative, *FP* false positive, *N* negative, *P* positive, *PN* predicted negative, *PP* predicted positive, *PTS* primary tumour staging, *T* total, *TN* true negative, *TP* true positive, *GPT-4* generative pre-trained transformer 4, *rPC* recurrent prostate cancer.
*Output set for GPT-4-only mentioning rPC differed differed significantly ($p < 0.05$) from the reviewed standard (McNemar's exact test).

Table 4 | Qualitative analysis of all six PTS reports not detected by GPT-4

PTS reports not detected by GPT-4			
Patient	**GPT-4 explanation — *translated from German to English***	**Root cause of LLM error**	**Expert-system check**
15	"The indication for the PET/CT examination seems to be more related to recurrence search, therapy, or follow-up monitoring due to the rapid PSA increase and already identified metastases, not primary tumour staging"	*Hallucinated evidence:* Pre-existing metastases were not mentioned in the report; LLM inferred them.	Table III shows the full expert-system-LLM dialog for this patient regarding PTS.
50	"The PET/CT indication is suspicion of tumour recurrence, not primary-tumour staging."	*Incomplete premises:* The LLM assumed that a PET/CT performed on a previously biopsied patient is a follow-up exam. However, biopsies followed shortly by a PET/CT indicate PTS.	Rule **PTS-2** with prostate was biopsied and rule **PTS-4** with PSA > 25 ng/ml.
80	"The patient already has a known prostate-cancer diagnosis and has received radical radiotherapy and hormone therapy, so the PET/CT is for follow-up, not PTS."	*Hallucinated evidence:* The report mentioned explicitly that no therapy was performed.	Rule **PTS-3** TNM with c-modifier and no "no PTS"→ keep PTS.
173	"The purpose was to check for bone or lymph-node metastases, suggesting follow-up or recurrence search, not PTS."	*Misinterpreted wording:* "question of metastasis" in diagnostic exam indication as proven metastasis.	Rule **PTS-5** compares diagnosis date with scan date (≤80 days) → still PTS possible.
187	"Extent evaluation before planned radiotherapy and suspected nodal metastasis indicate follow-up or therapy planning, not PTS."	*Misinterpreted wording:* "Suspected lymph node metastasis. Expansion prior to planned radiotherapy" in diagnostic exam indication as proven metastasis.	Rule **PTS-3** TNM with c-modifier→ keep PTS.
204	"Because metastases are mentioned and prostate cancer has already been diagnosed, the PET/CT appears to be for follow-up or metastasis search, not for primary-tumour staging."	*Hallucinated evidence:* no metastases were described and the patient history explicitly mentioned "no evidence for lymph node metastases". GPT-4 inferred their presence and re-labelled the study as recurrence follow-up.	Rule **PTS-3** TNM with c-modifier→ keep PTS.

The table lists each patient that GPT-4 missed, gives the original GPT-4 explanation translated to English, identifies (what we think) was the root cause, and shows how the Plato-3 expert system over-ruled the error. *GPT-4* generative pre-trained transformer 4, *LLM* large language model, *PET/CT* positron emission tomography combined with computed tomography, *PSA* prostate-specific antigene, *TNM* Classification of malignant tumours with tumour size (T), lymph nodes (N), and distant metastases (M).

beyond task-specific applications to an 'artificial expert' capable of handling a wider range of challenges.

Generic interfacing of fundamentally different AI and software types is nearly impossible without paradigm integration. Paradigm integration itself became possible because the required progress for understanding and supporting composition operations within programming environments[46] has been finally met. As a result, we have developed interfaces suitable for paradigm unification and the realization of the cognitive computing[34] and neuro-symbolic AI paradigm. Our approach mirrors real-world processes running at different abstraction levels[28], simplifying them into manageable components, while maintaining core functionalities of its individual modules[29]. This contrasts with contemporary multi-paradigm software, which use programming paradigms mostly in an isolated manner to address very specific problems.

Cognitive computing, which imparts meaning to data from context and intention, is ideal for understanding, designing, and controlling complex systems that handle heterogeneous data and exhibit unexpected behaviour[34,43]. While solving complex problems, such as compiling studies from unstructured medical data can be difficult, validating the problem's solution is generally easier. Therefore, our neuro-symbolic AI uses the LLM for problem solving but employs a rule-based expert system to control input and validate the LLM's outputs against its human designed ontology. Compared to simply using a knowledge base to augment an LLM[57], our expert system with its ontology uses recursive predicate logic[59] to verify and refine decisions[60]. This enables iterative reasoning, self-referencing[61], and the handling of emergent properties. Furthermore, creating an ontology effectively lessens the need to train specific AI foundation models and allows using boilerplate LLMs.

Contrary to standalone generative AI, our neuro-symbolic AI can be allowed to self-modify[62] and to acquire new knowledge from unstructured sources like medical textbooks or scientific papers while working on its tasks. We are currently exploring this capability to have GPT-4 write new rules into the ontology, while the expert system constructs thereof new prompts for GPT-4, solving the ontology-scaling problem[63].

Three caveats must be noted. First, the ontology was tailored to a small local data set, lacking rules for some parameters, such as pathology

localization and lesion quantity. However, Plato-3 recognizes and communicates unknowns, handling such cases by using fall-back rules. Second, GPT-4 receives continuous updates, so the results reflect the system's abilities at a particular time. The expert system's ontology, however, rectifies output regardless of LLM updates. Third, mistakes undetected by the neuro-symbolic AI in the original physician-generated reference would be carried over to the new reviewed reference, i.e. the new ground-truth. Also, some circular-validation risk remains whenever an AI may outperform its human benchmark. To curb that risk we accepted a revision only when the expert system's rule trace and the adjudicator's reasoning independently agreed.

Currently under development is a method for giving direct access to a workstation and using a multi-modal, vision-capable LLM, where our system can enter data directly into trial forms without changing an existing workflow. Using an expert system for auditing LLM decisions also offers the chance for LLM applications to pass medical device safety regulations, such as receiving a CE marking[64]. However, this does not absolve future work from including multi-centre data to test robustness across institutional reporting styles and to check against possible demographic bias introduced by the LLM.

Nevertheless, the lessons learned from this proof of principle are already applicable to large multicentre clinical trials: With our paradigm-integrating methods, language barriers, idiosyncrasies, or incompatible information technology no longer hinder evaluations of transnational datasets. Furthermore, LLMs under the transparent control of an expert system can be applied wherever humane rules and values must be respected[19], such as indexing electronic patient dossiers or supervising clinical workflows[5].

Conclusion

We conclude that our work offers a solution to errors and omissions, lack of transparency, and privacy risks of generative AI. Our rule-based quality control of LLMs permits their safe use structuring free-text from radiological reports. Hyper-exponential growth of AI technology will increasingly integrate AI into human-cantered tasks like health services, and future AI will likely instruct people rather than merely assisting them[4]. Having comprehensible AI decisions will therefore be crucial. Our work prepares for this paradigm shift by incorporating controlling and auditing mechanisms into autonomous AI systems, towards addressing the recognized needs for transparent, fair, robust, and ethical AI[40].

Data availability

The source data and statistical analyses underlying Table 1 and Table 3 are provided as Supplementary Data 1 (Excel file). All patient findings and reports can be shared in anonymized form upon reasonable request. All data requests must be submitted to author A.A.O (ali.afshar@insel.ch).

Code availability

The RUDS code that supports the findings of this study are available from SCCE GmbH but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of SCCE GmbH. The ontology code and GPT-4 prompts are given in the Supplementary Data 2.

Received: 5 February 2025; Accepted: 7 October 2025;
Published online: 21 November 2025

References

1. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **30**, 2613–2622 (2024).
2. AlSaad, R. et al. Multimodal large language models in health care: applications, challenges, and future outlook. *J. Med. Internet Res.* **26**, e59505 (2024).
3. Younis, H. A. et al. A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: applications, considerations, limitations, motivation and challenges. *Diagnostics* **14**, 109 (2024).
4. Park, P., Goldstein, S., O'Gara, A., Chen, M. & Hendrycks, D. AI deception: A survey of examples, risks, and potential solutions. *Patterns* **5**, 100988 (2024).
5. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, 141 (2023).
6. Chen, K., Xu, W. & Li, X. The potential of gemini and GPTs for structured report generation based on free-text 18F-FDG PET/CT Breast Cancer Reports. *Acad. Radiol.* **32**, 624–633 (2024).
7. Gumustop, S. et al. Predicting health crises from early warning signs in patient medical records. *Sci. Rep.* **12**, 19267 (2022).
8. Bagheri, A., Giachanou, A., Mosteiro, P. & Verberne, S. Natural language processing and text mining (turning unstructured data into structured). in *Clinical Applications of Artificial Intelligence in Real-World Data* (eds Asselbergs, F. W., Denaxas, S., Oberski, D. L. & Moore, J. H.) 69–93 (Springer International Publishing, Cham, 2023).
9. Achiam, O. J. et al. GPT-4 Technical Report. *OpenAI* preprint arXiv:2303.08774 (2023).
10. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
11. Ogden, C. K. & Richards, I. A. *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism* (Harcourt, Brace & World, Inc., New York, USA, 1923).
12. Liu, Y. et al. Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology* **1**, 100017 (2023).
13. Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N. Engl. J. Med.* **388**, 1233–1239 (2023).
14. Chakraborty, M. Explainable neural networks: achieving interpretability in neural models. *Arch. Comput. Methods Eng.* **31**, 3535–3550 (2024).
15. Liu, H. et al. Evaluating the logical reasoning ability of ChatGPT and GPT-4. Preprint at https://arxiv.org/abs/2304.03439 (2023).
16. D'Angelo, F. & Henning, C. On out-of-distribution detection with Bayesian neural networks. Preprint at https://arxiv.org/abs/2110.06020 (2021).
17. Nguyen, D., MacKenzie, A. & Kim, Y. H. Encouragement vs. liability: How prompt engineering influences ChatGPT-4's radiology exam performance. *Clin. Imaging* **115**, 110276 (2024).
18. Alberts, I. L. et al. Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? *Eur. J. Nucl. Med. Mol. Imaging* **50**, 1549–1552 (2023).
19. Yu, K. H., Healey, E., Leong, T. Y., Kohane, I. S. & Manrai, A. K. Medical artificial intelligence and human values. *N. Engl. J. Med*. **390**, 1895–1904 (2024).
20. Nau, D. S. Expert computer systems. *Computer* **16**, 63–85 (1983).
21. Hamilton, K., Nayak, A., Božić, B. & Longo, L. Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. *Semantic Web* **15**, 1265–1306 (2024).
22. Sotos, J. G. MYCIN and NEOMYCIN: two approaches to generating explanations in rule-based expert systems. *Aviat. Space Environ. Med.* **61**, 950–954 (1990).
23. Toosi, A., Bottino, A., Saboury, B., Siegel, E. & Rahmim, A. A brief history of AI: how to prevent another winter (a critical review). *PET Clin.* **16**, 449–469 (2021).
24. Clancey, W. J. The epistemology of a rule-based expert system —a framework for explanation. *Artif. Intell.* **20**, 215–251 (1983).
25. Patel, P. V. et al. Large language models outperform traditional natural language processing methods in extracting patient-reported outcomes in inflammatory bowel disease. *Gastro Hep Adv.* **4**, 100563 (2025).
26. Sheth, A., Roy, K. & Gaur, M. Neurosymbolic artificial intelligence (why, what, and how). *IEEE Intell. Syst.* **38**, 56–62 (2023).

27. Hamilton, K., Nayak, A., Božić, B. & Longo, L. Is neuro-symbolic AI meeting its promise in natural language processing? A structured review. *Semantic Web* **5**, 1265–1306 (2024).
28. Brenner, J. E. Levels of abstraction; levels of reality. in *Luciano Floridi's Philosophy of Technology: Critical Reflections* (ed Demir, H.) 201–222 (Springer Netherlands, Dordrecht, 2012).
29. Fishwick, P. A. The role of process abstraction in simulation. *IEEE Trans. Syst. Man Cybern.* **18**, 18–39 (1988).
30. Mämmelä, A., Riekki, J. & Kiviranta, M. LoosE Coupling: An Invisible Thread In The History Of Technology. *IEEE Access* **11**, 59456–59482 (2023).
31. Eder, J. & Schrefl, M. Coupling and cohesion in object-oriented systems. *Working paper, resp. Technical Report, University of Klagenfurt* (Institut für Wirtschaftsinformatik - Data & Knowledge Engineering, Altenberger Straße 69, 4040 Linz, Österreich, 1993).
32. Abramsky, S. & Sadrzadeh, M. Semantic unification. in *Categories and Types in Logic, Language, and Physics: Essays Dedicated to Jim Lambek on the Occasion of His 90th Birthday* (eds Casadio, C., Coecke, B., Moortgat, M. & Scott, P.) 1–13 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2014).
33. Van-Roy, P. & Haridi, S. *Concepts, Techniques, and Models of Computer Programming* (Prentice-Hall, 2004).
34. Bempong, B. O. The cognitive programming paradigm - the next programming structure. *Am. J. Softw. Eng. Appl.* **2**, 54–67 (2013).
35. Bhuyan, B. P., Ramdane-Cherif, A., Tomar, R. & Singh, T. P. Neuro-symbolic artificial intelligence: a survey. *Neural Comput. Appl.* **36**, 12809–12844 (2024).
36. Hiebert, J. & Lefevre, P. Conceptual and procedural knowledge in mathematics: an introductory analysis. in *Conceptual and Procedural Knowledge: The Case of Mathematics*. 1–27 (Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US, 1986).
37. Xu, C., Wang, Y. & Studer, T. A logic of knowing why. *Synthese* **198**, 1259–1285 (2021).
38. Glaudemans, A. W. J. M. et al. The first international network symposium on artificial intelligence and informatics in nuclear medicine: "The bright future of nuclear medicine is illuminated by artificial intelligence". *Eur. J. Nucl. Med. Mol. Imaging* **51**, 336–339 (2024).
39. Haug, C. J. & Drazen, J. M. Artificial intelligence and machine learning in clinical medicine, 2023. *N. Engl. J. Med.* **388**, 1201–1208 (2023).
40. Lekadir, K. et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* **388**, e081554 (2025).
41. Saboury, B. et al. Artificial intelligence in nuclear medicine: opportunities, challenges, and responsibilities toward a trustworthy ecosystem. *J. Nucl. Med.* **64**, 188–196 (2023).
42. Afshar-Oromieh, A. et al. Performance of [68Ga]Ga-PSMA-11 PET/CT in patients with recurrent prostate cancer after prostatectomy—a multi-centre evaluation of 2533 patients. *Eur. J. Nucl. Med. Mol. Imaging* **48**, 2925–2934 (2021).
43. Moto-oka, T. Overview to the fifth generation computer system project. *SIGARCH Comput. Archit. News* **11**, 417–422 (1983).
44. *Federal Act on Research involving Human Beings (Human Research Act, HRA)* (Swiss Federal Law AS 2013 3215, 2011).
45. Turner, R. Computational abstraction. *Entropy* **23**, 213 (2021).
46. Zave, P. A compositional approach to multiparadigm programming. *IEEE Softw.* **6**, 15–25 (1989).
47. Sperber, D. & Wilson, D. *Relevance: Communication and Cognition* (Wiley-Blackwell, 1995).
48. Nii, H. P. Blackboard application systems, blackboard systems and a knowledge engineering perspective. *AI Mag.* **7**, 82 (1986).
49. Graci, R. Understanding the significance of situational context and common ground in communication. in *Aphasia's Implications for Linguistics Research: Exploring the Interface Between Semantics and Pragmatics* 27–51 (Springer International Publishing, Cham, 2023).
50. Kampis, G. Self-modifying systems. in *Biology and Cognitive Science: A New Framework for Dynamics, Information, and Complexxity* (Pergamon Press, Oxford, UK, 1991).
51. McKinney, E. H. & Yoos, C. J. Information about information: a taxonomy of views. *MIS Q.* **34**, 329–344 (2010).
52. Wooldridge, M. & Jennings, N. R. Intelligent agents: theory and practice. *Knowl. Eng. Rev.* **10**, 115–152 (1995).
53. Coplien, J. O. & Reenskaug, T. M. H. The data, context and interaction paradigm. In *Proc. 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity* 227–228 (Association for Computing Machinery, Tucson, Arizona, USA, 2012).
54. Djuric, D. & Devedzic, V. Incorporating the ontology paradigm into software engineering: enhancing domain-driven programming in Clojure/Java. *IEEE Trans. Syst. Man, Cybern. Part C.* **42**, 3–14 (2012).
55. Fernando, C. N. P. & David, H. D. W. Definite clause grammars for language analysis—A survey of the formalism and a comparison with augmented transition networks. *Artif. Intell.* **13**, 231–278 (1980).
56. Chomsky, N. Three models for the description of language. *IRE Trans. Inf. Theory* **2**, 113–124 (1956).
57. Li, Y. et al. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus* **15**, e40895 (2023).
58. Kang, T., Turfah, A., Kim, J., Perotte, A. & Weng, C. A neuro-symbolic method for understanding free-text medical evidence. *J. Am. Med. Inform. Assoc.* **28**, 1703–1711 (2021).
59. Kleene, S. C. Recursive predicates and quantifiers. *Trans. Am. Math. Soc.* **53**, 41–73 (1943).
60. McCabe, H. Formal Logic For Expert Systems. in *AI and Cognitive Science '92* (eds Ryan, K. & Sutcliffe, R. F. E.) 334–337 (Springer London, London, 1993).
61. Kauffman, L. Self-reference and recursive forms. *J. Soc. Biol. Syst.* **10**, 53–72 (1987).
62. Fishwick, P. A., Narayanan, N. H., Sticklen, J. & Bonarini, A. A multimodel approach to reasoning and simulation. *IEEE Trans. Syst. Man Cybern.* **24**, 1433–1449 (1994).
63. Potter, S. A. Survey of Knowledge Acquisition from Natural Language. *Artificial Intelligence Applications Institute, Division of Informatics, University of Edinburgh* https://www.aiai.ed.ac.uk/project/akt/work/stephenp/TMA%20of%20KAfromNL.pdf (2001).
64. European Parliament and Council of the European Union. Regulation (EU) 2017/745 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. *Official Journal of the European Union* **L 117**, 1–175 (2017).

Acknowledgements

This study was funded by the department of nuclear medicine, Inselspital Bern. Zentit GmbH and SCCE GmbH contributed in-kind by providing the RUDS software and labour.

Author contributions

G.A.P. designed and led the study, developed the RUDS software, engineered the expert system ontology, analysed the data, performed the logic concordance check, prepared figures and tables, and wrote the manuscript. T.K.W. contributed to RUDS development, created Plato-3, engineered the expert system ontology, designed the ontology figure, and wrote the manuscript. S.C.B. contributed to RUDS development and assisted with manuscript preparation and writing. C.M. manually extracted study parameters from reports for comparison with the AI systems. G.M. served as study nurse, managed patient data, and ensured compliance with ethics requirements. L.P.M. assisted with manuscript preparation and study design. K.Y.S. provided technology support. A.R. contributed patient data and domain knowledge for the ontology. A.A.O. managed ethics

requirements, contributed patient data, coordinated with the study nurse, provided ontology domain knowledge, and manually extracted study parameters from reports for comparison with the AI systems. Authors G.A.P. and T.K.W. contributed equally.

Competing interests

G.A.P. is co-founder and CEO of Zentit GmbH. T.K.W. is founder and CEO of SCCE GmbH. Both companies have been founded for commercialization of RUDS and Plato-3. S.C.B. was an employee of Zentit GmbH. All other authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Open-AI's GPT-4o in order to shorten certain sentences for staying within the word count. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-025-01194-x.

Correspondence and requests for materials should be addressed to George A. Prenosil.

Peer review information *Communications Medicine* thanks Sneha Mithun, Anirban Mukhopadhyay and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.