Computer Science Review 60 (2026) 100902

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Neuro-symbolic agentic AI: Architectures, integration patterns, applications, open challenges and future research directions

Safayat Bin Hakim [a], Muhammad Adil [b,1,*], Alvaro Velasquez [c], Houbing Herbert Song [a]

[a] Dept. of Information Systems, University of Maryland, Baltimore County, Baltimore, MD, USA
[b] Dept. of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA
[c] Dept. of Computer Science, University of Colorado Boulder, Boulder, CO, USA

ARTICLE INFO

Keywords:
Neuro-symbolic AI
AI agents
Symbolic reasoning
Neural networks
Explainability
Meta-cognition
Learning and inference
Knowledge representation
Logic and reasoning
Trustworthiness

ABSTRACT

Neuro-symbolic AI synergizes neural networks' pattern recognition with symbolic reasoning's logical structure, addressing fundamental limitations each paradigm exhibits independently. This systematic survey analyzes 178 papers (2020–November 2025) using PRISMA methodology, establishing a comprehensive taxonomy across architectural configurations (single-agent, multi-agent) and integration dimensions: knowledge representation (44%), learning and inference (63%), logic and reasoning (35%), explainability and trustworthiness (28%), and meta-cognition (5%). We identify critical research imbalances, meta-cognitive capabilities remain severely underexplored despite demonstrating greater performance impact than sophisticated integration patterns alone. Through architectural analysis spanning sequential, parallel, end-to-end differentiable, and unified representation approaches, we examine prominent systems (Agent Q, GoalAct, AlphaGeometry, Reflexion, MetaGPT) and evaluate their effectiveness across robotics, natural language processing, autonomous vehicles, healthcare, and education domains. Performance comparisons reveal consistent neuro-symbolic superiority: 23% improvement in robotic task completion, 95.4% autonomous navigation success versus 18.6% neural baselines, and order-of-magnitude reductions in sample complexity. We expose persistent challenges—reproducibility barriers, weak generalization, scalability constraints, symbol grounding difficulties—and propose structured solutions through TRAP-inspired meta-cognitive frameworks, standardized evaluation protocols, and hierarchical agentic architectures balancing symbolic decomposition with neural adaptability.

1. Introduction

AI has experienced cyclical growth and decline—AI summers and winters [1]. Currently in the third AI summer, we witness significant deep learning advances and renewed interest in integrating neural approaches with symbolic reasoning [2,3]. AI agents—autonomous software programs [4]—perceive environments, reason, plan, and execute actions effectively [5–8]. Recent advances (2024–2025) enable agents to coordinate in multi-agent environments, maintain long-term context through narrative memory, and employ refined chain-of-thought prompting with external tools [9–11].

Despite impressive capabilities, neural and symbolic approaches face fundamental limitations independently. Neural methods excel at pattern recognition and adaptation but lack interpretability, struggle with systematic reasoning, and require extensive training data [12,13]. Symbolic approaches offer explicit rule-based reasoning and transparency but demonstrate brittleness in uncertain environments and limited learning capabilities [14]. These complementary strengths suggest a natural synergy through neuro-symbolic AI.

Neuro-symbolic AI agents combine neural networks' flexibility with symbolic reasoning's logical structure and interpretability, creating systems that learn from data while maintaining explainable decisions. This integration addresses key questions: How can neuro-symbolic agents ground abstract symbolic representations in real-world sensory data? How can neuro-symbolic systems improve generalization to novel tasks beyond training distributions? What strategies ensure transparent, interpretable reasoning processes? How can neuro-symbolic models learn robustly from small or sparse datasets? These questions motivate our comprehensive analysis of architectures, integration patterns, and deployment strategies that enable effective neural-symbolic synergy.

Our main contributions are: (1) The first comprehensive survey on neuro-symbolic AI agents, reviewing 178 papers (2020–November

* Corresponding author.
Email address: muhammad.adil@ieee.org (M. Adil).
[1] Present address: Dept. of Computer Science, Texas Southern University, Houston, TX, USA.

https://doi.org/10.1016/j.cosrev.2026.100902
Received 19 May 2025; Received in revised form 27 November 2025; Accepted 12 January 2026
Available online 31 January 2026

2025) using rigorous PRISMA methodology; (2) A novel taxonomy encompassing architectural configurations (single vs. multi-agent, vertical vs. horizontal) and integration strategies (sequential, parallel, end-to-end differentiable, unified representation); (3) Systematic performance analysis demonstrating neuro-symbolic advantages across diverse domains with quantitative comparisons; (4) Identification of critical research gaps, particularly meta-cognition (5%) and explainability (28%); (5) Structured open challenges and future directions addressing benchmarking consistency, integration strategies, and meta-cognitive capabilities via frameworks like TRAP.

1.1. Survey methodology

We employed the PRISMA methodology [15] ensuring a comprehensive, and rigorous review. Fig. 1 illustrates our systematic selection process maintaining breadth and depth while capturing emerging neuro-symbolic AI agent trends.

1.1.1. Search strategy

We queried IEEE Xplore, ACM Digital Library, Springer Link, arXiv, Google Scholar, and ScienceDirect, combining "neurosymbolic" OR "neuro-symbolic" with agent-related terms ("agent", "multi-agent", "autonomous system") and integration approaches ("reasoning", "learning", "knowledge representation", "explainability", "meta-cognition"). The initial search yielded 1428 papers from January 2020 to November 2025.

1.1.2. Inclusion and exclusion criteria

We focused on English-language papers including peer-reviewed journal articles, conference papers, book chapters, and highly-cited arXiv preprints containing relevant advances. While we prioritize peer-reviewed publications, we selectively include influential arXiv preprints that demonstrate significant architectural innovations, have catalyzed follow-on research, or provide unique empirical insights not yet available in the peer-reviewed literature (e.g., Agent Q's MCTS-guided web navigation [16], GITM's text-based Minecraft agent [17]). We acknowledge that such preprints have not undergone formal peer review and may contain limitations identified during review processes; notably, several included preprints have been submitted to top-tier AI venues (e.g., NeurIPS, ICLR, AAAI) with acceptance rates below 25%, have incorporated reviewer feedback in revised arXiv versions, or are currently under review in subsequent submission cycles. Papers must explicitly address both neuro-symbolic integration and agent architectures, providing empirical results or theoretical frameworks. We excluded papers focusing exclusively on neural or symbolic approaches without integration, those not addressing agent-based systems, surveys without original contributions, and works lacking sufficient technical detail.

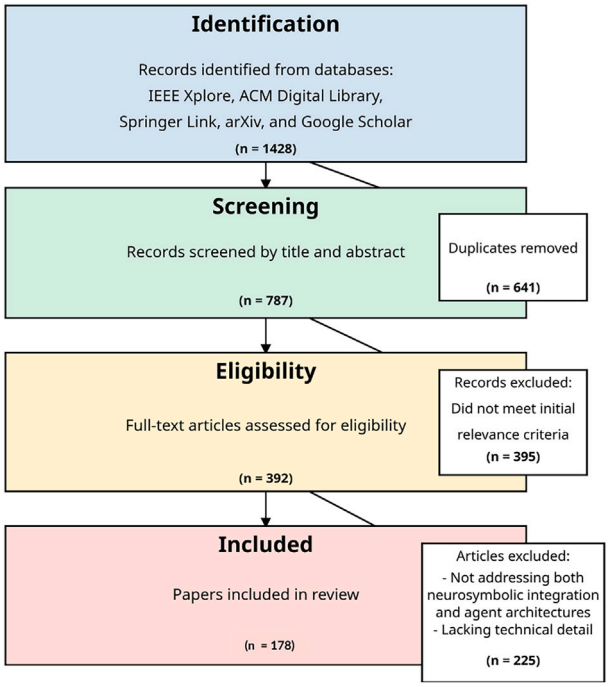

Fig. 1. PRISMA flow diagram for paper selection, identifying 1428 initial records, ultimately including 178 papers meeting all neuro-symbolic AI agent research criteria (2020–November 2025).

1.1.3. Selection process

After removing 641 duplicates, 787 papers underwent title/abstract screening, yielding 392 candidates. Full-text review reduced this to 178 papers meeting all criteria, as shown in Fig. 1.

1.1.4. Data extraction and analysis

From each paper, we extracted bibliographic information, system architecture, primary research focus, application domains, evaluation methods, limitations, reproducibility status, and identified failure modes. Papers were categorized according to our taxonomy and analyzed quantitatively and qualitatively to identify trends, gaps, and limitations.

1.2. Paper organization

Section 2 provides background on AI agents, symbolic AI, neural approaches, and neuro-symbolic integration. Section 3 presents our comprehensive taxonomy. Section 5 analyzes architectural components and prominent systems. Sections 4 through 11 present applications, methodologies, performance analysis, and evaluation methods. Section 10 explores meta-cognition. Section 12 presents the AlphaGeometry case study. Sections 13–15 discuss open challenges, future directions, and conclusions.

2. Background and definitions

This section familiarizes readers with the integration of neural and symbolic components within AI agent architectures, outlining key characteristics and limitations grounding their integration.

AI agents are autonomous software entities perceiving environments, making decisions, and taking actions to achieve goals [8]. Key characteristics include autonomy (operating without intervention), reactivity (responding to changes), proactivity (goal-directed behavior), and social ability (interacting with agents/humans) [18]. Agent evolution progressed from simple reflex agents acting on current percepts to learning agents adapting through experience. Recent advances have produced sophisticated agents based on large language models [6], multimodal systems [7,19], and embodied agents [20] capable of complex planning and reasoning, as illustrated in Fig. 2.

2.1. Symbolic AI

Symbolic AI represents knowledge through explicit symbols and rules [21], founded on the Physical Symbol System Hypothesis positing human cognition models through symbol manipulation [22,23]. Key characteristics include explicit knowledge representation, logical reasoning, transparency, and modularity. While symbolic systems excel in interpretability and structured reasoning, they struggle with uncertainty, unstructured data, and adapting to novel situations. Traditional symbolic approaches continue offering advantages in constrained environments prioritizing computational efficiency, formal verification, or explicit rule-based reasoning [24–26].

2.2. Neural approaches

Neural approaches derive representations implicitly from data, excelling in pattern recognition and adaptation. Characterized by data-driven learning, distributed information representation across network

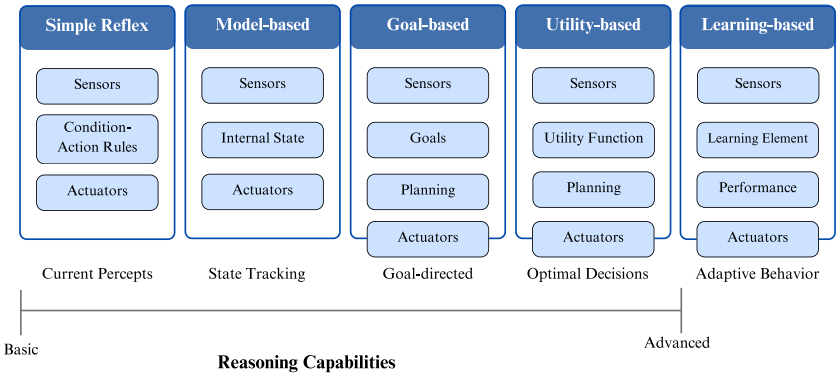

Fig. 2. Evolution of AI agent architectures from simple reflex agents to learning agents with increased reasoning capabilities.

weights, adaptability to varying conditions, and powerful pattern recognition [27,28]. Despite their strengths, neural systems face challenges in systematic reasoning, explicit knowledge representation, and decision-making interpretability, which are particularly problematic in domains requiring formal logical inference or transparent justification.

2.3. Neuro-symbolic integration

Neuro-symbolic AI merges neural and symbolic paradigm strengths [29], creating hybrid systems that combine neural network flexibility and pattern recognition with symbolic reasoning's logical structure and interpretability. Integration is motivated by dual-process reasoning theories [30,31] distinguishing System 1 (intuitive) versus System 2 (deliberative) thinking [32], supported by recent meta-analyses [33] and computational models [34]. Integrated learning cycles combine neural pattern recognition with symbolic knowledge representation [29], while modular design patterns [35] provide computational frameworks for implementing dual processes. Cognitive architectures like Soar [36,37] and ACT-R [38] demonstrate the realization of hybrid processing.

2.4. Agentic AI: recent advances

Recent research (2024–2025) extended traditional neuro-symbolic approaches developing agentic AI frameworks—autonomous systems proactively designing workflows and coordinating multiple agents for complex tasks. Innovations like narrative memory for long-term contextual reasoning [39,40] enable agents to maintain coherent understanding across temporally distant events. Refined chain-of-thought prompting improved agents' capacity to utilize external tools and coordinate workflows. These developments underscore neuro-symbolic integration's critical role in enabling truly agentic behavior, as neither purely neural nor symbolic approaches easily achieve this combination of flexibility and structured persistence.

3. Taxonomy of neurosymbolic AI agents

Based on systematic literature analysis, we propose a comprehensive taxonomy considering both architectural configurations and integration approaches, as illustrated in Fig. 3.

3.1. Architectural configurations

Neuro-symbolic agent architectures are broadly categorized as single-agent or multi-agent systems. Single-agent systems integrate neural and symbolic components for independent task performance, representing 65% of reviewed papers. These are subdivided by integration patterns: *Sequential coupling* operates modules in sequence, as in ReAct [41] and Reflexion [42], providing clear separation but risking information loss at boundaries [43]. *Parallel coupling* processes inputs concurrently, combining outputs for rapid pattern recognition and structured reasoning, handling uncertainty well but potentially introducing conflicts. *End-to-end differentiable architectures* embed symbolic operations within differentiable networks, facilitating gradient-based optimization, as in Logic Tensor Networks [44,45] and Neural Theorem Provers [46]. *Zero-shot concept learning systems* like ZeroC [47] enable novel concept recognition through symbolic graph integration with neural energy-based models.

Table 1
Key capabilities across major neuro-symbolic agent frameworks.

Framework	Auton.	Meta-cog.	Tool Int.	Symb. Reason.	Neural Learn.	Explainability
Agent Q [16]	✓	✗	✓	✓	✓	✓
GoalAct [48]	✓	✗	✓	✓	✓	✗
AlphaGeometry [49]	✗	✗	✗	✓	✓	✓
Reflexion [42]	✗	✓	✗	✗	✓	✓
MetaGPT [50]	✓	✗	✓	✗	✓	✓
TRAP [51]	✓	✓	✗	✓	✓	✓

Table 1 summarizes key capabilities across major neuro-symbolic frameworks, showing varying support for autonomy, meta-cognition, tool integration, symbolic reasoning, neural learning, and explainability.

Multi-agent architectures [52–54], comprising 35% of the literature, address complex tasks through collaboration [55]. *Vertical architectures* implement hierarchical structures with leader agents coordinating subordinates, improving task efficiency [56]. *Horizontal architectures* operate as egalitarian systems with peer collaboration, exemplified by AgentVerse [56] and MetaGPT [50]. *Hybrid architectures* combine vertical and horizontal elements with dynamic leadership. Emerging dynamics emphasize dynamic task allocation and enhanced collaborative reasoning [57,58] adapting to changing conditions.

3.2. Integration approaches

Neuro-symbolic integration divides into four main approaches. *Knowledge Representation Integration* [59] leverages neural enhancements to symbolic knowledge bases, organizing neural representations per symbolic schemas [60–64]. Recent advances in symbolic knowledge distillation [65] enable the extraction of structured, interpretable symbolic knowledge from LLMs, complementing knowledge graph integration by transforming implicit neural representations into explicit symbolic forms that enhance transparency and reasoning capabilities. *Learning and Inference Integration* embeds symbolic reasoning within neural frameworks, employing differentiable reasoning [66–69], neural-guided symbolic search (AlphaGeometry [49]), neuro-symbolic program synthesis [70], and continual learning [71,72]. LINC [73] demonstrates significant improvements in logical reasoning by integrating formal logic with neural language understanding.

Explainability and Trustworthiness Integration derives transparent symbolic explanations from neural decisions [74,75], incorporating

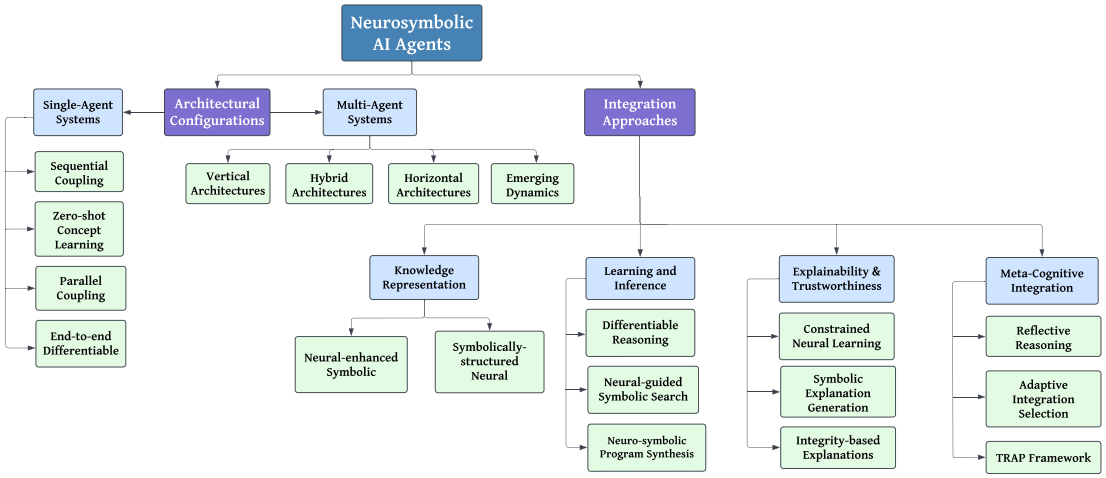

Fig. 3. Comprehensive taxonomy of neuro-symbolic AI agents, considering architectural configurations and integration approaches.

symbolic constraints during training [76,77]. The CREST framework [78,79] fosters trust through interpretable processes and bias mitigation [80]. *Meta-Cognitive Integration* enables self-monitoring and adaptive reasoning, though it is severely underexplored (5%).

Our analysis shows Learning and Inference integration dominates (63%), followed by Knowledge Representation (44%) and Logic and Reasoning (35%). Explainability and Trustworthiness (28%) and Meta-Cognitive integration (5%) remain underexplored, as visualized in Fig. 5. Alternative classification schemes complement our integration-focused framework. Kautz's influential taxonomy [1] organizes neuro-symbolic systems by the level at which neural and symbolic components interact—from symbolic knowledge compilation into neural networks to neural guidance of symbolic reasoning—providing an orthogonal perspective on integration depth and directionality that enriches understanding of neuro-symbolic architectures.

3.3. Conceptual overview of integration patterns

Neuro-symbolic integration patterns enable agentic capabilities combining neural flexibility with symbolic structure. Sequential integration supports autonomy linking neural perception to symbolic action planning. Parallel integration enhances proactivity processing inputs concurrently through neural and symbolic pathways. End-to-end differentiable architectures facilitate goal-directed learning, making symbolic operations trainable through gradient-based optimization. Unified representation approaches like ZeroC support generalization mapping symbolic graphs to neural energies [47], and recognizing novel concepts by composing existing ones. Detailed mathematical formulations appear in Appendix A.

Table 2 summarizes the taxonomy dimensions for neuro-symbolic AI agents.

Table 2
Summary of taxonomy dimensions for neuro-symbolic AI agents.

Dimension	Categories/ Sub-categories	Representative examples/ References
Architectural Configurations	Single-Agent	Sequential (ReAct [41], Reflexion [42]), Parallel, End-to-End Differentiable (Logic Tensor Networks [44]), Zero-shot (ZeroC [47])
	Multi-Agent	Vertical [56], Horizontal (AgentVerse [56], MetaGPT [50]), Hybrid, Emerging Dynamics [57,58]
Integration Approaches	Knowledge Representation	Neural-enhanced symbolic knowledge (SymAgent [59]), knowledge graphs [60], concept learners [61]
	Learning & Inference	Differentiable reasoning [66,67], Neural-guided search (AlphaGeometry [49]), LINC [73]
	Explainability & Trust	Symbolic explanations [74], Constrained learning [76], CREST [78]
	Meta-Cognitive	Reflexion [42], TRAP [51], Adaptive selection [81]

4. Applications and use cases

This section analyzes key application domains showing how neuro-symbolic systems address limitations in reasoning, knowledge integration, and explainability. Table 3 provides a granular comparison of neuro-symbolic systems against traditional baselines, detailing the specific architectural mechanisms and quantitative performance shifts observed across these domains.

4.1. Robotics and embodied AI

Robotic systems require perceptual understanding of unstructured environments and structured reasoning for task planning. Neuro-symbolic approaches enable robots to interpret ambiguous instructions, decompose high-level goals, and adapt plans when encountering obstacles [84]. Studies show approximately 23% higher success rates in complex manipulation tasks compared to purely neural or symbolic methods [85]. Despite progress, most systems are evaluated in controlled environments with limited variability. The symbol grounding problem—connecting abstract symbols to real-world perceptions—remains challenging, particularly with novel objects [86].

4.2. Natural language processing

NLP systems must bridge unstructured language with structured reasoning, maintain factual consistency, and provide transparent explanations. ReAct [41] demonstrates how neuro-symbolic integration reduces hallucination rates (6% vs. 14%) and improves multi-hop reasoning accuracy by combining neural language understanding with structured reasoning traces. LINC [73] achieves 75.3% ± 2.1% accuracy on the FOLIO benchmark, compared to 74.8% for pure neural approaches, by combining neural language models with symbolic logical verification. However, even LINC struggles with disordered premises and distractors, suggesting brittleness when facing input variability.

Table 3
Analysis of neuro-symbolic performance gains. Metrics are grounded in specific experimental contexts (e.g., human studies, oracle guidance) to ensure accurate comparison.

Domain	Task & benchmark	Architecture	Mechanism of gain	Performance shift
Robotics	Embodied reasoning (ALFWorld simulator)	**ReAct** (Reasoning + Acting)	Sparse "thoughts" decompose goals and track subgoals over long horizons [41]	**Success Rate (All):** 37% (BUTLER, IL) → **71%** (ReAct, best of 6) [41]
NLP	Multi-hop QA (HotpotQA)	**ReAct** (+ Wikipedia API)	External retrieval reduces unsupported reasoning traces and grounds facts [41]	**Hallucinated Reasoning:** 14% (CoT) → **6%** (ReAct, human study of 50 cases) [41]
Healthcare	Diabetes diagnosis (Pima Indians Dataset)	**LNN** (Logical Neural Net)	Learnable weights on logical rules capture multi-factor risk pathways [82]	**Accuracy (Test Split):** 76.95% (Random Forest) → **80.52%** (LNN) [82]
Education	Advanced math (15 AIME problems)	**NEOLAF** (Neuro-symbolic agent)	"System 2" symbolic solvers guided by "System 1" LLM planning [83]	**Accuracy (PoC Study):** 7% (GPT-4) → **80%** (NEOLAF with Oracle) [83]
Web Ops	Online shopping (WebShop)	**ReAct** (Sparse Reasoning)	Reasoning bridges noisy page observations and action selection [41]	**Success Rate:** 30.1% (Act-only) → **40.0%** (ReAct) [41]

4.3. Proactive reasoning in neuro-symbolic agents

Proactive reasoning—anticipating future needs and acting in advance—is essential for agentic behavior. Chain-of-Thought (CoT) prompting [87,88] guides neural models in generating intermediate reasoning steps, effectively injecting symbolic scaffolds. Tree of Thoughts [89] enables systematic reasoning branch exploration, while Self-Consistency [90] leverages ensemble generation. Internal simulations and world models enhance proactive reasoning by incorporating mental environment representations that anticipate outcomes. Emerging applications of world models extend to telecommunications, where agentic systems employ generative state-space reasoning for network management [91], demonstrating the generalizability of proactive reasoning paradigms across diverse domains. Metacognitive loops provide another mechanism by periodically evaluating progress against goals. These loops typically implement symbolic verification routines that check neural output coherence against predefined constraints, aligning with "slow thinking" strategies [92,93].

4.4. Decision support systems

In healthcare, neuro-symbolic systems reduce diagnostic errors while providing interpretable explanations thereby increasing clinician trust. The CREST framework [78] enhances trustworthiness through consistency, reliability, explainability, and safety integration, incorporating clinical guidelines as symbolic constraints guiding neural pattern recognition. Despite these advantages, clinical simulation benchmarks like AgentClinic reveal challenges including limited simulation fidelity, subjective evaluation metrics, and a lack of longitudinal data integration [94].

4.5. Autonomous vehicles

Autonomous driving requires rapid perception and guaranteed safety properties. Perception systems combine neural object detection with symbolic scene understanding. Planning frameworks integrate neural trajectory prediction with rule-based safety verification, enabling flexible traffic adaptation while ensuring compliance. Recent work demonstrates a 43% safety violation reduction compared to purely neural methods [51]. However, systems face challenges in extreme scenarios not considered during design, with symbolic components relying on hand-crafted rules that do not cover all edge cases.

4.6. Education and training

Educational AI must adapt to individual learning patterns while maintaining pedagogical structure. Neuro-symbolic tutoring agents personalize experiences through: modeling student knowledge using symbolic knowledge graphs; adapting teaching strategies based on neural predictions; providing structured feedback connecting performance to specific concepts. Studies show significantly higher learning gains with neuro-symbolic tutoring [83,95,96]. However, effectiveness depends heavily on the quality of knowledge representation, often requiring domain expert input.

5. Neuro-symbolic agent architectures and components

This section analyzes how different neuro-symbolic architectural approaches compare in performance and address limitations in reasoning, knowledge integration, and explainability. Building on our taxonomy, we examine key architectural components and prominent systems.

5.1. Symbolic task planners and planning-focused architectures

Symbolic task planners decompose high-level goals into actionable subtasks, enhancing goal-directedness. Techniques like PDDL [97,98] and hierarchical task networks [99] provide structured planning frameworks, as in AutoGPT+P [100], supporting robotic task execution. GoalAct employs hierarchical planning requiring thoughtful domain expansion. While symbolic components provide structure, they may require domain-specific adaptation limiting generalizability. Nevertheless, symbolic planning integration with neural execution achieves approximately a 23% improvement in robotic task completion [85].

5.2. Neural semantic parsers and language interpretation

Neural semantic parsers bridge natural language and symbolic representations, supporting proactive reasoning. Toolformer [101] enables agents to self-supervise external symbolic tool use, grounding language in executable actions. LINC achieves 75.3% ± 2.1% accuracy on the FOLIO benchmark using StarCoder+, compared to 74.8% for pure

neural approaches [73]. However, even LINC struggles with linguistic complexities, performing poorly with distractors or disorder. These limitations align with broader findings on transformer compositionality limits, where agents face "capability cliffs" on complex multi-step tasks [102].

5.3. Hierarchical planning and goal management

Goal-directedness is fundamental to agentic AI. GoalAct [48] exemplifies neuro-symbolic integration for goal-directed behavior through its LLM-based framework explicitly maintaining global plans and decomposing tasks hierarchically. GoalAct continuously updates the symbolic global goal state representation guiding the skill hierarchy, and reducing planning complexity through symbolic decomposition combined with neural execution. LegalAgentBench evaluations demonstrate a 95.06% success rate versus 93.33% for Plan-and-Solve [48].

Hierarchical Language Agent (HLA) [103] implements a cognitive architecture splitting processing into "Slow Mind" (powerful LLM) for high-level intent reasoning, "Fast Mind" (lightweight model) for generating macro-actions, and "Executor" (symbolic/reactive policy) for low-level actions. In cooperative games, HLA agents achieved approximately 50% higher scores than the baselines.

Ghost in the Minecraft (GITM) [17] demonstrates symbolic goal decomposition power in open-world environments. GITM became the first agent to obtain all 262 items in Minecraft's tech tree, achieving 95% success on the ObtainDiamond task versus baselines managing only 10%, 3%, and 20% [17]. However, GITM operates in a text-based simulation avoiding visual perception challenges, which raises questions about its performance in the actual game environment.

5.4. Neuro-symbolic executors and action-focused architectures

Neuro-symbolic executors implement environmental actions, combining neural perception with symbolic reasoning. The ReAct framework [41] alternates between reasoning steps and action execution in structured cycles, as shown in Fig. 4. ReWoo [104] offers an alternative by decoupling reasoning from observation.

Reflexion [42] extends this architecture incorporating structured self-reflection. However, reliance on LLM self-critique can lead to hallucinated feedback, particularly in long-horizon tasks. Its sliding window memory approach limits the ability to leverage insights from earlier experiences [105].

5.5. Autonomy in neuro-symbolic agents

Agent Q [16] demonstrates how neural-guided symbolic search dramatically enhances autonomy through Monte Carlo Tree Search (MCTS) combined with neural self-critique and reinforcement learning. OpenTable evaluations show Agent Q boosted performance from 18.6% (zero-shot) to 81.7% after training, further to 95.4% with MCTS integration [16]. Despite impressive results, Agent Q struggles with generalizing to novel web tasks unseen during training. AgentGen [106] addresses scalability by automating diverse planning environment generation.

StarPO framework [107] addresses training instability by optimizing complete decision trajectories, mitigating "Echo Trap" where agents are stuck in self-repeating loops. AlphaEvolve [108] pushes autonomy toward recursive self-improvement through evolutionary coding with agents autonomously discovering algorithms.

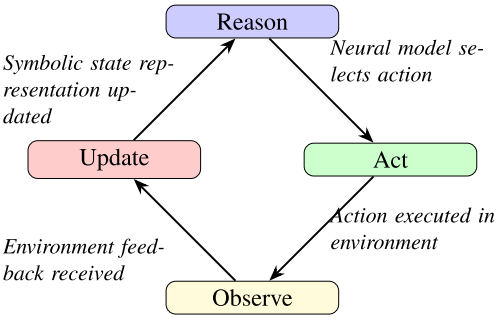

Fig. 4. Simplified ReAct cycle integrating symbolic reasoning and neural adaptability for iterative agentic behavior.

5.6. Memory architectures for agentic AI

Memory integration combines short-term neural scratchpads with long-term symbolic memory structures. GITM [17] incorporates a symbolic memory buffer storing successful plans retrievable and adaptable for new tasks, storing up to N = 5 successful action sequences summarized into canonical plans. Recent narrative memory integration [39,40] has enhanced context maintenance through structured mechanisms organizing episodic information. Retrieval-Augmented Generation (RAG) techniques [109] and their associated benchmarks [110] further enhance memory access.

Effective memory architectures differentiate between: *Episodic memory* recording specific events with symbolic structure [111]; *Semantic memory* storing general knowledge in knowledge graphs; *Working memory* functioning as a short-term scratchpad. By integrating these memory types, neuro-symbolic agents maintain coherence over extended interactions.

5.7. Zero-shot concept learning architectures

ZeroC [47] implements a unified neuro-symbolic framework dramatically improving generalization. This architecture represents concepts as symbolic graphs mapping to neural energy-based models. Key innovation: new concepts are acquired and recognized at inference time by composing existing concepts according to symbolic descriptions, without additional training. This enables exceptional generalization correctly identifying novel visual concepts based solely on symbolic descriptions.

5.8. Multi-agent neuro-symbolic frameworks

Multi-agent neuro-symbolic frameworks organize specialized agents collaborating on complex tasks [53,54]. AgentVerse [56] implements a horizontal architecture with a four-stage workflow: recruitment, collaborative decision-making, independent action execution, and evaluation. MetaGPT [50] addresses communication inefficiencies through structured information sharing via a shared environment.

AutoGen [112] enables structured agent collaboration through conversational protocols [55], allowing multiple LLM-based agents to coordinate on complex tasks. The CAMEL framework [113] leverages role-playing and "inception prompting" facilitating autonomous cooperation. MetaGPT integrates Standard Operating Procedures (SOPs) into agent prompts, guiding specialized agents to collaborate through structured processes. CulturePark [114] demonstrated how symbolic personas shape emergent agent behaviors in cross-cultural interactions.

5.8.1. Social coordination and ethical challenges

SOTOPIA [115] highlights challenges in multi-agent coordination and ethical alignment. LLM-based agents underperform humans in complex social scenarios, revealing gaps that compromise performance and interpretability. Symbolic interaction protocols provide a promising solution by defining social norms that constrain neural agents [116]. The BNE-Q framework ensures ethical behavior by validating actions against predefined norms. These approaches underscore the neuro-symbolic framework needs managing social dynamics and ethical risks.

Effective collaboration relies on structured mechanisms integrating neural and symbolic components through message passing with defined information access rights [50,56,117].

5.9. Formal decision-making frameworks

Formal decision-making frameworks provide mathematical foundations for agent behavior under uncertainty. The BNE-Q framework [118] combines Bayesian Nash Equilibrium concepts with multi-agent LLM reasoning, modeling collaborative reasoning as Dec-POMDP. Through iterative optimization, the system converges to consistent joint decisions satisfying equilibrium constraints. This demonstrates complementary neuro-symbolic integration advantages: neural components provide flexible reasoning while symbolic Bayesian networks provide formal convergence guarantees.

In centralized training with decentralized execution (CTDE) approaches, LLM-based coordination addresses scalability challenges [119]. These systems use large language models to coordinate agents by providing shared plans, effectively reducing the need to learn everything from scratch.

5.10. Multimodal neuro-symbolic agents

Large Multimodal Agents (LMAs) [7] extend architectures handling diverse input modalities (vision, audio, touch). These systems integrate neural perception modules for processing multimodal data with symbolic reasoning frameworks for planning and action execution. While increasing complexity, this enables agents to operate in dynamic, multimodal environments more closely resembling the real world.

5.11. Integration patterns and their effectiveness

Table 4 summarizes the strengths and limitations of major integration patterns. Sequential patterns represent the most common approach (58%), benefiting from conceptual clarity but facing information loss at integration boundaries. Parallel integration demonstrates the strongest performance in uncertain environments but comes with the highest computational requirements. End-to-end differentiable approaches offer unified gradient-based optimization but face substantial practical limitations when scaling to complex reasoning tasks due to computational complexity and semantic preservation challenges. Unified representation approaches like ZeroC demonstrate exceptional generalization but remain limited to specific domains.

As Table 5 shows, different integration patterns exhibit varying capabilities across key performance dimensions.

Meta-cognitive integration, represented by the TRAP framework, offers the most comprehensive capabilities across all dimensions. This approach adaptively selects and combines integration patterns based on task requirements, enabling systems to leverage the appropriate integration strategy for different task aspects.

Table 4
Comparison of integration patterns in neuro-symbolic agent architectures.

Integration pattern	Strengths	Limitations
Sequential Neural-to-Symbolic	Strong in perception-to-reasoning; clear separation	Information loss at boundaries; limited feedback
Sequential Symbolic-to-Neural	Effective for reasoning-to-generation; logical consistency	Rigid planning; limited adaptability
Parallel Integration	Robust uncertain environment handling; complementary strengths	Complex integration; potential conflicts
End-to-End Differentiable	Trainable optimization; gradient-based learning	Computational complexity; limited symbolic expressivity
Unified Representation	Strong generalization; one-to-one mapping	Design complexity; domain specificity

Table 5
Integration patterns vs. key outcomes.

Pattern	Expl. ability	Samp. Eff.	Gen.	Realtime Adapt.	Multi-Ag. Support
Sequential	✓	✗	✗	✓	✗
Parallel	✓	✗	✓	✓	✓
End-to-End Diff.	✗	✓	✓	✗	✗
Unified Rep.	✓	✓	✓	✗	✗
Meta-Cog. (TRAP)	✓	✓	✓	✓	✓

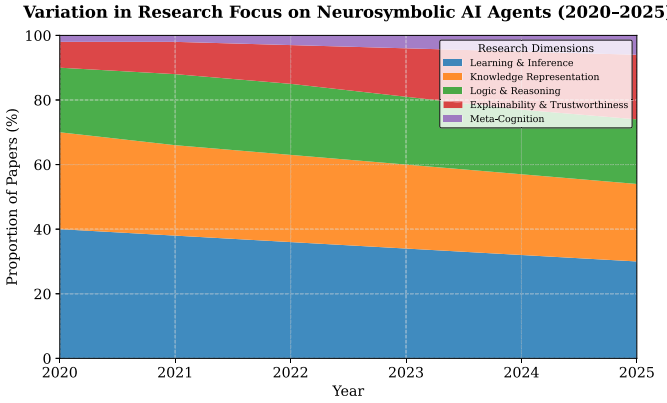

Fig. 5. Temporal evolution of neuro-symbolic AI agent research focus (2020–November 2025) across five dimensions. Percentages represent proportion of 178 papers addressing each dimension (non-exclusive categories): learning & inference (63%), knowledge representation (44%), logic & reasoning (35%), explainability & trustworthiness (28%), and meta-cognition (5%).

6. Results and analysis

This section analyzes the current state of neuro-symbolic AI agent research, identifying key patterns and evaluating different approaches' comparative performance.

6.1. Research trends

Our 178 paper analysis revealed significant variation in research focus from 2020 to November 2025, as Fig. 5 illustrates. Learning and inference mechanisms receive the most attention (63%), reflecting the field's strong focus on developing effective computational models. Knowledge representation follows as the second most studied area (44%), highlighting the importance of creating hybrid representations. Logic and reasoning account for 35%. Explainability and trustworthiness appear in only 28%, revealing a significant gap between technical capabilities and human-centered considerations. Most strikingly, meta-cognition appears in just 5%, making it the most underexplored dimension despite its potential importance for creating truly autonomous systems.

Notable temporal patterns include exponential increase in neuro-symbolic AI agent research since 2020, reflecting growing recognition of complementary strengths. Growing attention to explainability since 2022 is likely driven by increasing awareness of these properties' importance for real-world deployment. Meta-cognition research focus emergence only in 2023 suggests emerging awareness but significant room.

Intersection analysis revealed some area combinations as well-explored while others remain largely untapped. Knowledge representation and learning integration appear in 27%. However, knowledge

Table 6
Performance comparison of neurosymbolic vs. pure approaches.

Task domain	Pure neural	Pure symbolic	Neurosymbolic
Geometry Theorem (IMO-AG-30)	–	10/30 problems	25/30 problems [49]
Question Answering (HotpotQA [120] EM)	29.4%	25.7%	35.1% [41]
MNIST Addition (N = 1)	97.3%	–	97.9% [13]
Logical Reasoning (FOLIO)	74.8%	–	75.3% [73]
Robotic Planning (Success Rate)	67.3%	72.8%	83.2% [85]
Goal Management (Success Rate)	–	–	+12.22% [48]
Autonomous Web Navigation	18.6%	–	95.4% [16]
Multi-Agent Cooperation Score	–	–	+50% [103]
Hallucination Rate (HotpotQA)	14%	–	6% [41]
Memory Efficiency (Plan Reuse)	N/A	N/A	5 retrieval plans [17]
Sample Efficiency (to 90% Acc.)	10,000+	N/A	100–1000 [61]
Reasoning Transparency (Rating)	1.8/5	4.2/5	4.0/5 [79]

representation and explainability are addressed together in only 4%, highlighting a critical gap. Notably, only a single paper (AlphaGeometry [49]) addresses all four main research areas, indicating significant opportunities for more holistic approaches.

6.2. Performance comparison

Comparative analysis of neuro-symbolic agent performance across different architectural approaches indicates significant advantages over purely neural or symbolic approaches, as Table 6 summarizes.

Our performance analysis revealed key findings demonstrating neuro-symbolic integration advantages. First, neuro-symbolic approaches consistently outperform both pure neural and pure symbolic methods across diverse tasks. This pattern holds across domains requiring different capabilities, suggesting integration benefits reflect the fundamental complementarity between neural and symbolic processing. Second, the performance advantage is most pronounced in tasks requiring both pattern recognition and structured reasoning, such as geometry theorem proving and logical reasoning.

Agentic capabilities show particularly dramatic improvements, demonstrated by Agent Q's autonomous web navigation performance improving from 18.6% to 95.4% with neuro-symbolic MCTS integration [16]. HLA demonstrated 50% higher cooperation scores in multi-agent tasks [103]. Memory efficiency shows significant gains through structured symbolic representations, as seen in GITM's plan reuse capabilities [17].

In human interaction terms, neuro-symbolic methods achieve comparable transparency ratings to pure symbolic approaches (4.0/5 vs. 4.2/5) while maintaining neural methods' flexibility. Sample efficiency shows perhaps the most dramatic improvement, with some systems requiring 1–2 orders of magnitude fewer examples than pure neural approaches.

Important to note, these performance improvements often come with trade-offs in computational efficiency and system complexity. AlphaGeometry solved 25/30 IMO problems representing a dramatic performance improvement [121], yet faces scaling challenges. Authors explicitly note "scaling beyond Olympiad geometry remains an open question" [49].

6.3. Architecture effectiveness analysis

Different architectural approaches demonstrate varying effectiveness across task types. Single-agent sequential architectures demonstrate strong performance in well-defined tasks with clear perception-to-action mappings. However, these may struggle with information loss at integration boundaries. Multi-agent approaches excel in complex tasks requiring diverse capabilities and parallel processing [57]. Vertical architectures show efficiency advantages in hierarchical task decomposition. Horizontal architectures demonstrate stronger collaborative problem-solving in open-ended domains, as shown in AutoGen [112] and MetaGPT [50].

Hierarchical planning systems like GoalAct [48] and HLA [103] improve goal management through structured decomposition. Zero-shot learning architectures like ZeroC [47] demonstrate exceptional generalization but remain domain-limited. Fast-slow cognitive architectures like Plan-SOFAI [122] exhibit superior adaptability across varying problem difficulties.

Integration pattern effectiveness varies by task domain. Sequential patterns demonstrate strong performance in structured tasks. Parallel integration shows robust uncertain environment handling but with increased computational requirements. End-to-end differentiable approaches face practical scaling limitations. These varying effectiveness profiles highlight the importance of matching architectural choices to application requirements.

6.4. Lessons learned

Our systematic analysis uncovers significant patterns synthesizing insights across domains. Integration pattern effectiveness varies substantially by application domain rather than exhibiting uniform advantages. This challenges the notion of "best" integration approach, suggesting architectural choices should be guided by specific requirements.

Striking imbalance exists between research attention on single-agent architectures (65%) compared to multi-agent approaches (35%), despite the latter showing stronger performance on complex collaborative tasks. This disproportion likely reflects additional coordination challenges, but our analysis suggests that performance improvements justify greater research focus.

Contrary to prevailing assumptions, explicit meta-cognitive capabilities prove more valuable than sophisticated integration patterns in long-horizon tasks. Systems with even simple self-monitoring mechanisms consistently outperform more complex approaches lacking these capabilities. This suggests future developments should prioritize meta-cognitive integration alongside advances in core neural-symbolic processing techniques.

6.5. Research gaps

Our analysis identified several critical research gaps limiting progress. Severe meta-cognitive research underrepresentation [123] (5%) indicates a major opportunity for advancing self-monitoring and adaptive reasoning capabilities. Few systems implement comprehensive self-assessment, strategy adaptation, or reflective reasoning essential for autonomous operation [51]. While frameworks like TRAP [51,124] provide structured approaches, their limited adoption indicates significant implementation gaps. This gap is particularly concerning given our finding that meta-cognitive capabilities often contribute more to complex task performance than sophisticated integration patterns.

Despite theoretical transparency advantages, explainability research remains limited (28%). Recent systematic reviews [14,43] indicate that most neuro-symbolic systems show only medium-low explainability. The intersection of explainability with other research areas remains especially sparse, with only 4% addressing both explainability and knowledge representation.

While various integration approaches exist, standardized methodologies for effectively combining neural and symbolic components are lacking [125]. Few papers address integration quality formal verification or provide metrics assessing semantic consistency. This methodological gap makes systematically comparing different integration approaches difficult.

Standardized benchmarks specifically designed for neuro-symbolic integration are limited [126]. Metrics assessing integration quality, cross-component knowledge transfer, and adaptive capabilities remain underdeveloped. This evaluation gap complicates progress assessment and makes fair system comparison difficult, as shown in Table 7.

Methods for scaling neuro-symbolic systems to handle large knowledge bases while maintaining computational efficiency represent another gap [127]. Few papers address the combinatorial explosion in reasoning paths or provide efficient updating mechanisms for integrated representations.

Insufficient work connects formal agentic properties (autonomy, goal-directedness, proactivity) with neuro-symbolic integration. While recent advances in Agent Q [16], GoalAct [48], and BNE-Q [118] demonstrate potential, systematic research remains sparse. Addressing these gaps requires interdisciplinary approaches combining insights from cognitive science, formal methods, and human-computer interaction.

7. Practical considerations for deployment

7.1. Computational efficiency and scalability

Computational efficiency represents critical neuro-symbolic AI deployment consideration, as systems must balance neural inferential demands with symbolic logical processing. Recent analyses demonstrate precipitous inference cost declines for large language models, with GPT-3.5-equivalent systems experiencing 280-fold cost reduction from $20 to $0.07 per million tokens over 18 months, representing annual reductions of 9–900× depending on task complexity [128]. However, advanced reasoning paradigms introduce substantial overhead: Tree of Thoughts requires 5–100× more generated tokens than Chain-of-Thought [89], while ReAct necessitates multiple iterative API calls per decision [41]. Emergent research into “slow thinking” reasoning [92] demonstrates compute-optimal scaling strategies improving test-time efficiency by more than 4×, with smaller models using adaptive inference-time computation outperforming 14× larger models [93].

Neurosymbolic architectures combining neural networks with symbolic search exhibit distinct computational profiles. AlphaGo employed 40 search threads with 48 CPUs and 8 GPUs for single-machine deployment, while competitive configurations utilized 1202 CPUs and 176 GPUs [129]. Training costs reached unprecedented scales, with GPT-4 requiring an estimated $78 million and Gemini Ultra consuming approximately $191 million [130]. A recent systematic review analyzing 167 papers identifies computational complexity as the primary deployment challenge [14]. Asynchronous execution strategies partially mitigate costs by parallelizing neural evaluations on GPUs while conducting symbolic search on CPUs. Empirical evaluations demonstrate hybrid approaches achieve superhuman performance while evaluating substantially fewer positions—AlphaGo evaluated thousands of times fewer positions than Deep Blue by selecting positions intelligently through neural policy networks [129]. These trade-offs suggest neuro-symbolic architectures are most advantageous where interpretability, logical consistency, and sample efficiency outweigh increased computational costs.

8. Challenges and limitations

This section examines how current neuro-symbolic approaches attempt to overcome limitations while identifying remaining challenges. Despite significant advances, neuro-symbolic AI agents face substantial hurdles requiring attention.

8.1. Scalability challenges

Knowledge integration scalability remains a fundamental obstacle as systems expand to handle larger, more complex domains. These systems face several critical challenges: *Knowledge Integration Complexity* escalates poorly as knowledge bases expand, and managing contradictions between learned neural patterns and symbolic constraints becomes increasingly difficult [131]. Current approaches typically limit the scope to manageable subdomains or implement hierarchical knowledge structures, but these strategies often trade generality for tractability. *Computational Resource Requirements* are substantial, particularly when combining large neural models with complex symbolic reasoning. End-to-end training requires significant memory for gradient computation across hybrid architectures [45]. *Deployment Constraints* present particular challenges for on-device deployment [132], limiting practical applications in settings like mobile devices with limited computational resources.

8.2. Integration complexity

Representation alignment between neural and symbolic components represents a persistent challenge. This fundamental mismatch creates difficulties in establishing meaningful correspondences across the neural-symbolic boundary [125]. Current approaches include neural-symbolic embeddings [44], concept bottleneck models [133], and neuro-symbolic concept learners [61], each with specific strengths and limitations. However, none fully resolves the challenge of creating seamless integration, particularly for complex knowledge structures with many interconnected concepts.

Training and optimization challenges further complicate neuro-symbolic integration. Symbolic components often introduce non-differentiable operations complicating gradient-based learning, and creating discontinuities that standard neural training methods struggle to navigate. End-to-end optimization requires bridging fundamentally different learning paradigms within a unified framework [13]. Techniques such as relaxed logical operations, neural theorem proving [134], and reinforcement learning with symbolic rewards [135] provide partial solutions but require further development to address the full range of integration challenges.

8.3. Ethical and privacy concerns

Integrating neural and symbolic components creates complex interactions raising significant ethical and privacy concerns. *Bias Integration Complexity* arises as neural components may learn biases from training data, while symbolic rules can either enforce fairness constraints or encode problematic assumptions [78]. Interaction between these sources creates dynamics that are difficult to analyze and mitigate. *Emergent Behavior Risks* manifest in systems like CulturePark’s cultural dialogues and SOTOPIA’s social interactions, which may develop unintended biases through agent interactions. Neurosymbolic solutions, such as symbolic intent verification through BNE-Q’s Bayesian constraints or explicit ethical rule sets, can constrain neural outputs to align with human values. *Privacy Vulnerabilities* are heightened as symbolic components may make sensitive information more explicit than purely neural representations. Integration across domains can lead to unexpected information leakage through symbolic representations explicitly capturing sensitive relationships [43].

8.4. Human-agent collaboration challenges

Neurosymbolic approaches present unique challenges and opportunities for human-agent collaboration directly affecting practical utility. A fundamental issue in collaborative settings is “automation surprise” problem, where human users are confused by unexpected agent behavior. By maintaining symbolic joint activity representations, neuro-symbolic agents can detect potential misalignments between plans and human expectations, then use neural components to generate appropriate explanations. This capability is particularly valuable in high-stakes domains where teams must maintain shared situational awareness. For example, in healthcare decision support, symbolic diagnostic reasoning verification identifies when agent conclusions might contradict clinician expectations, prompting detailed explanation generation. This transparency enhances trust and enables more effective collaboration.

Neural and symbolic components’ complementary strengths create promising collaboration opportunities. Symbolic components provide explicit shared goals, plans, and constraints representations, making agent reasoning transparent. Meanwhile, neural components adapt to

individual human preferences, communication styles, and expertise levels, creating more natural interaction experiences. This combination addresses a fundamental challenge: balancing predictability with adaptability creating systems that are both reliable and responsive to human needs.

Different integration patterns show varying effectiveness for collaborative tasks. Sequential integration might offer clearer agent reasoning explanations, while parallel integration could better handle uncertain or ambiguous human inputs. Evaluating these trade-offs across domains with different transparency versus flexibility requirements remains an important research direction.

8.5. Evaluation and benchmarking limitations

Current neuro-symbolic agent evaluation approaches face significant limitations hampering progress assessment. Existing benchmarks often focus exclusively on either neural or symbolic capabilities, rather than their integration [126], making assessing neuro-symbolic approaches' added value difficult. Comparative evaluation is complicated by varying assumptions, while standardized integration quality assessment metrics remain undeveloped [136]. Furthermore, evaluating long-term adaptability requires extensive resources that are often unavailable in research settings.

Common benchmarking frameworks exhibit specific weaknesses limiting the utility of neuro-symbolic evaluation. AgentBench lacks multi-agent coordination and long-term memory assessment capabilities, with methodology critiques raising concerns about prompt sensitivity [136]. The FOLIO benchmark has a limited scope compared to human reasoning capabilities and is prone to memorization rather than genuine deduction [73]. WildBench features English-only tasks limiting cross-linguistic evaluation, and faces reproducibility issues due to proprietary components, and annotation inconsistencies [137]. AgentClinic shows concerns regarding simulation fidelity and lacks longitudinal data essential for healthcare applications [94]. StrategyQA focuses primarily on answer accuracy rather than the validity of the reasoning process [138].

Recent efforts in developing neuro-symbolic-specific benchmarks [94,137] represent important steps, but require further refinement and broader adoption. Developing targeted metrics for integration quality, cross-component knowledge transfer, and adaptive capabilities would significantly enhance evaluation ability.

8.6. Symbol grounding problem

Symbol grounding problem, connecting abstract symbols with real-world referents via perception or sensorimotor interaction [86,139], directly addresses the challenge of grounding symbolic representations in sensory data. How can we bridge the gap between neural perception processing continuous, noisy sensory data and symbolic reasoning operating on discrete representations? This issue is particularly critical for neuro-symbolic agents in embodied contexts such as robotics and autonomous vehicles.

As Fig. 6 illustrates, the challenge involves bridging continuous perceptual inputs and discrete symbolic representations. This fundamental issue—critically reviewed by Taddeo and Floridi [140]—requires methods for learning symbol meanings without extensive supervision [86]. Situated grounding through environmental interaction represents a promising approach allowing agents to learn symbol meanings through experience [141], enabling more robust connections that generalize to novel situations. Contemporary frameworks increasingly address this through modular design patterns explicitly structuring the interface between neural perception and symbolic reasoning [35].

8.7. Reproducibility challenges

Neurosymbolic AI agent research faces significant reproducibility challenges that limit independent verification and slow collective progress. Systematic evaluation of the 25 most prominent neuro-symbolic agent systems (2019-2025) revealed a bifurcated reproducibility landscape strongly correlated with system type and institutional backing: while framework-oriented works demonstrate strong code availability (18 of 25 systems, 72%) with public repositories and basic documentation, only 6 systems (24%) provide pre-trained model weights, 15 systems (60%) lack explicit hardware requirement documentation, and 19 systems (76%) depend on external commercial APIs whose non-deterministic updates fundamentally compromise long-term parameter-level reproducibility. Even systems with exemplary open-source practices face substantial barriers—AlphaGeometry [49] provides complete code and weights yet requires computational resources to generate nearly 1 billion synthetic training premises, while Reflexion incurs prohibitive API costs through high token usage when using commercial language models. Industrial frameworks [49,50,112,115] exemplify emerging best practices through Docker containerization, hash-verified dependencies, and comprehensive artifact management, yet frontier systems [101,108] and theoretical contributions [51] (3 systems with zero code release) represent reproducibility extremes where verification is either results-only or entirely conceptual. The dominant reliance on prompt engineering over weight training has fundamentally shifted the reproducibility paradigm from "load state dict" to "replay interaction logs," necessitating new standards for behavioral reproducibility including frozen prompt templates, logged tool calls, explicit random seed specifications, and documented computational costs—infrastructure gaps that currently prevent independent verification across 48% of surveyed systems despite nominal code availability. MetaGPT's requirement for manual SOP design when adapting workflows and AutoGen's complex dependencies (which necessitated AutoGen Studio's development [112]) further illustrate deployment challenges that complicate reproduction even when code exists. The research community should prioritize comprehensive artifact releases with containerized environments, standardized benchmarking protocols, and transparent documentation of computational requirements to accelerate progress through effective knowledge sharing and collaborative development.

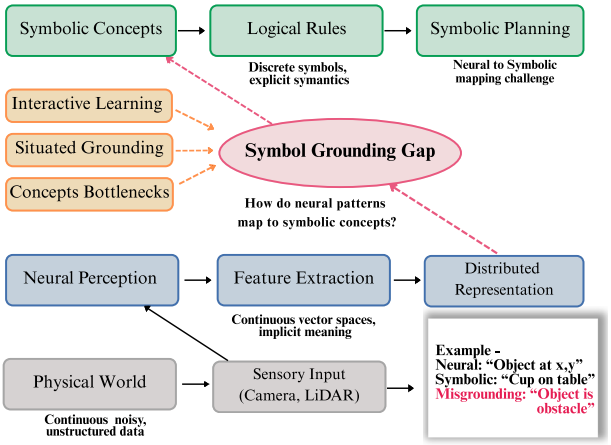

Fig. 6. Symbol grounding problem in neuro-symbolic agents, showing challenge mapping neural perceptions to symbolic representations with solutions like situated grounding.

9. Explainability and trustworthiness

This section examines how neuro-symbolic approaches enhance transparency and interpretability while maintaining neural flexibility. Explainability and trustworthiness represent critical yet underexplored dimensions [43,125], accounting for only 28% of surveyed literature despite the practical importance.

9.1. Explainability and trustworthiness mechanisms

Neurosymbolic agents implement various approaches enhancing explainability and trustworthiness, leveraging symbolic transparency while maintaining neural adaptability. *Symbolic reasoning traces* provide step-by-step decision process explanations, creating interpretable reasoning path records. ReAct [41] maintains reasoning chains as explanations, while logic-based approaches provide formal proof structures justifying conclusions. These processes address "black box" pure neural approaches.

Concept-based explanations ground decisions in human-unders tandable concepts, bridging neural processing and human comprehension. Neuro-symbolic concept learners [61] explain decisions using learned symbolic concepts corresponding to human-interpretable categories, while compositional explanations decompose complex decisions into simpler components [74]. By connecting neural processing to symbolic concepts, these approaches generate explanations that align with human understanding.

Verification and oversight mechanisms ensure system reliability through explicit checking procedures. Symbolic verification ensures adherence to critical constraints through logical consistency checking, formal proof generation, and safety property verification, while human-in-the-loop oversight provides plan verification interfaces and interactive correction mechanisms. These approaches demonstrate improved alignment with human values.

Integrity-based approaches [79] focus on fostering appropriate trust considering factors such as consistency, competence, and benevolence. The CREST framework [78] integrates these considerations generating explanations tailored to user needs and domain requirements, enhancing trustworthiness by providing the specific information users need for evaluating system reliability.

9.2. Limitations in current approaches

Despite advances, neuro-symbolic systems face significant explainability challenges limiting practical high-stakes domain adoption. A fundamental tension exists between optimizing for task performance and providing transparent explanations, particularly in complex domains. Representational gaps between neural representations and symbolic explanations can lead to post-hoc rationalizations rather than faithful accounts of the actual decision-making process, potentially undermining trust.

As neuro-symbolic systems tackle increasingly complex tasks, explanation complexity grows substantially, potentially overwhelming users with excessive detail. Finding the right abstraction level that balances completeness with comprehensibility remains challenging, requiring careful consideration of user expertise, time constraints, and specific information needs [125]. Furthermore, evaluation difficulties complicate progress assessment by relying heavily on subjective human judgments of explanation quality. The field lacks standardized metrics for evaluating explanation fidelity, utility, and user-friendliness, thereby slowing progress.

9.3. Research opportunities

Significant opportunities remain in advancing neuro-symbolic system explainability and trustworthiness. Developing standardized metrics for evaluating explanation quality across different architectures would enable a more systematic approach to comparison [126]. These metrics could address multiple explanation dimensions, including fidelity to the decision process, target user comprehensibility, and specific task usefulness. Investigating the relationship between explainability and actual trustworthiness through controlled user studies represents another important direction.

Creating scalable approaches for maintaining explainability in complex reasoning processes with deep integration chains would address current system critical limitations. Research on abstraction mechanisms, hierarchical explanations, and interactive exploration interfaces could help manage complexity while preserving interpretability. Exploring domain-specific explanation requirements [142] would enable more effective specific application explanation design. Developing frameworks that balance explanation comprehensiveness with cognitive load would help address the tension between completeness and usability.

Addressing these opportunities could significantly advance neuro-symbolic system practical deployment in high-stakes domains where explainability and trustworthiness are paramount. By developing more effective explanation approaches, researchers enhance neuro-symbolic system adoption and impact across applications where transparency is essential for user acceptance and regulatory compliance.

10. Meta-cognition in neurosymbolic AI agents

Meta-cognition—the ability to monitor, evaluate, and adaptively control cognitive processes—enhances neuro-symbolic system reasoning [123]. Despite being crucial for human-like intelligence [143], it remains severely underexplored, appearing in only 5% of reviewed papers [14], even though Section 6.4 findings show that meta-cognitive capabilities often contribute more to performance than sophisticated integration patterns. This section examines how meta-cognitive capabilities significantly enhance neuro-symbolic agents' adaptability, robustness, and autonomy.

10.1. Theoretical foundations

Meta-cognition in neuro-symbolic AI draws from cognitive science perspectives distinguishing between object-level cognition (direct task processing) and meta-level cognition (monitoring and controlling object-level processes). Nelson and Narens' framework [144] distinguishes metacognitive monitoring (awareness of knowledge, reasoning processes, limitations) from metacognitive control (directing cognitive resources based on monitoring insights), while Ackerman and Thompson [145] specialize this as *meta-reasoning*—a critical function monitoring thoughts determining strategy selection and termination.

Common Model of Cognition (CMC) [146] integrates cognitive architectures like ACT-R [38], Soar [36], and Sigma, incorporating metacognitive functions through working memory systems (maintaining current processing awareness), procedural memory for strategy selection (enabling adaptive control), declarative memory storing metacognitive knowledge (about strategies and effectiveness), and learning mechanisms modifying these structures based on experience. This model provides a comprehensive blueprint for implementing meta-cognitive capabilities by specifying memory structures and processes needed for effective self-monitoring and adaptation.

10.2. The TRAP framework for meta-cognition

TRAP framework [51], illustrated in Fig. 7, provides a structured meta-cognition approach in neuro-symbolic AI agents, focusing on four complementary aspects: Transparency, Reasoning, Adaptation, and Perception. *Transparency* involves explicitly representing and explaining the agent internal state and reasoning processes, achieved through function $g(f(x), \theta)$, enabling both self-monitoring and communication with human collaborators. *Reasoning* encompasses applying logical inference and evaluating the agent's own performance, represented by $f(x; g(\theta))$, where metacognitive AI incorporates self-reflection into logical processing. *Adaptation* represents the ability to modify strategies based on self-evaluation, formally captured as $f'(x; g(f(x), \theta))$, where f' is the adapted model based on metacognitive assessment. *Perception* enables detecting environmental or task context changes requiring strategy adjustments, represented by $f(g(x), x)$, allowing responsive adaptation to changing conditions.

TRAP framework emphasizes synergy between neural and symbolic components in meta-cognitive processes, where symbolic components provide explicit self-monitoring representations while neural components enable flexible adaptation across diverse contexts. This integration

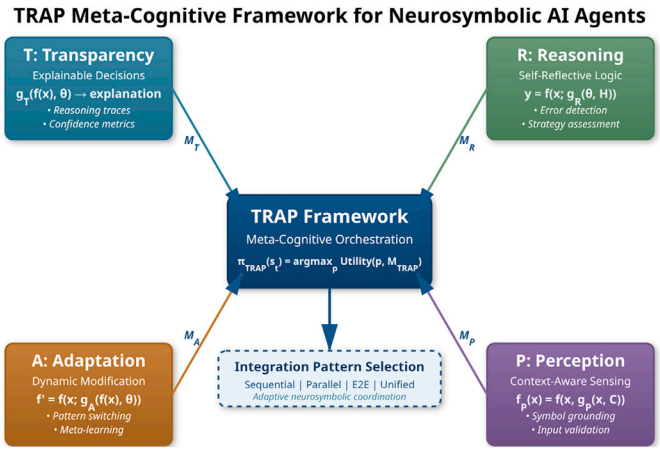

Fig. 7. The TRAP Meta-Cognitive Framework for Neurosymbolic AI Agents, illustrating integration of Transparency, Reasoning, Adaptation, and Perception components to orchestrate adaptive neuro-symbolic patterns.

occurs through mechanisms like abductive learning for adaptation and logic-tensor networks for Transparency, creating a comprehensive system combining complementary paradigm strengths.

10.3. Current approaches and mathematical formulation

Current meta-cognition approaches in neuro-symbolic agents include reinforcement learning for meta-cognitive policy learning, cognitive architecture integration for structured metacognitive processes, and reflective reasoning for self-criticism and adaptation. The Reflexion framework [42] exemplifies the reflective reasoning approach through a system with three LLMs serving different functions: Actor (generates actions), Evaluator (scores performance), and Self-Reflection model (generates reflections stored in memory). The framework maintains a sliding-window memory of past reflections prepended to prompts for future tasks [42], creating a persistent insight record informing future processing.

These meta-cognitive approaches can be formalized using a two-level framework distinguishing between object-level and meta-level processes:

$$\begin{aligned} M &= \text{Monitor}(S, O, f_O) \\ f'_O &= \text{Control}(f_O, M) \end{aligned} \tag{1}$$

where S represents state space, O the observation space, and $f_O : S \times O \rightarrow A$ maps states and observations to actions. Here, M represents metacognitive monitoring information, and f'_O represents the modified object-level function after metacognitive control. This formulation captures the hierarchical relationship between monitoring and control processes enabling self-regulation.

In neuro-symbolic implementations, these functions combine neural and symbolic components:

$$\begin{aligned} M &= f^M_{\text{neural}}(S, O, f_O) \oplus f^M_{\text{symbolic}}(S, O, f_O) \\ f'_O &= g(f_O, M) \end{aligned} \tag{2}$$

where $\oplus$ represents integration combining neural and symbolic monitoring, and g represents a function modifying the object-level process based on metacognitive insights. Additional mathematical formulations for TRAP-specific components appear in Appendix B.

10.4. Fundamental challenges in self-assessment

Fundamental difficulty in developing effective LLM-based agent meta-cognition lies in the bootstrapping problem [147] which limits self-assessment reliability. Many frameworks rely on the LLM itself to perform core meta-cognitive functions: assessing outputs, reflecting on errors, estimating confidence, or generating corrective actions [105]. Yet LLMs possess weaknesses in these very areas, exhibiting tendencies, flawed reasoning, poor calibration, and unreliable self-assessment [105]. This creates a circular dependency where potentially unreliable components construct mechanisms intended to enhance reliability.

This situation suggests purely LLM-driven internal meta-cognition might face inherent limitations due to circular reliance. Achieving robust, trustworthy self-correction likely necessitates grounding the meta-cognitive process in external verification sources—calls to external tools, checks against formal symbolic knowledge bases, validation against real-world feedback, or human oversight integration—rather than relying solely on LLM internal self-assessment. These external grounding mechanisms provide independent verification breaking self-assessment circularity, and creating more reliable meta-cognitive capabilities.

10.5. Research opportunities in meta-cognition

Limited meta-cognition exploration in neuro-symbolic agents indicates significant research opportunities. Developing comprehensive frameworks integrating neural and symbolic metacognitive processes across different time scales would enable more sophisticated self-monitoring and adaptation, combining rapid neural assessment with symbolic reasoning about longer-term patterns. Creating standardized evaluation metrics assessing metacognitive capabilities [148] and their overall system performance impact would facilitate more systematic research progress, addressing different meta-cognition aspects including monitoring accuracy, adaptation effectiveness, and resource allocation efficiency.

Exploring scalable approaches to metacognitive monitoring in complex reasoning tasks with multiple integration points would address the current approach key limitation. As reasoning processes grow more complex, maintaining effective monitoring becomes increasingly challenging. Research on abstraction mechanisms, hierarchical monitoring, and focused attention could help manage complexity while maintaining effective oversight. Investigating methods for transferring metacognitive knowledge across domains would enhance generalization capabilities, reducing the need for domain-specific metacognitive learning and enabling more rapid adaptation to new domains based on accumulated experience.

Designing techniques explaining metacognitive decisions to humans would enhance adaptive system trust by making self-monitoring and adaptation processes transparent. These explanations help users understand why the system changed its approach or identified potential issues, enhancing human and adaptive AI system collaboration. Addressing these opportunities could significantly advance neuro-symbolic agents' adaptability, robustness, and autonomy, enabling more effective performance in complex, dynamic environments.

11. Benchmarking and evaluation

This section examines current benchmarking approaches and proposes improved evaluation methodologies for neuro-symbolic agent systems. Effective evaluation requires specialized benchmarks and metrics assessing both individual capabilities and integrated performance, beyond traditional evaluations focusing on either neural or symbolic capabilities in isolation.

11.1. Current benchmarking approaches

Current neuro-symbolic agent evaluation approaches include task-specific benchmarks assessing particular capabilities (question answering, knowledge graph completion, theorem proving) and integration-focused evaluations specifically addressing neuro-symbolic integration. While task-specific benchmarks often focus on either neural or symbolic capabilities independently, integration-focused benchmarks like

AgentBench [136], WildBench [137], and AgentClinic [94] better assess combined capabilities but remain limited in comprehensively evaluating the full neuro-symbolic integration spectrum.

Recent additions include benchmarks evaluating agentic capabilities, such as SOTOPIA [115], specifically testing social intelligence and goal-directed behavior. Addressing the need for deeper diagnostic granularity, RSBench [149] systematically evaluates reasoning shortcuts using the `countrss` algorithm verifying concept quality under distributional shifts. Its empirical evaluation across arithmetic and logic domains demonstrates that obtaining high-quality concepts remains a significant challenge for both neural and neuro-symbolic architectures. Similarly, addressing trustworthiness in retrieval-augmented generation, LibreEval [150] provides a multilingual dataset of over 70,000 examples specifically designed for scalable hallucination detection. By employing a consensus labeling approach with multiple LLM judges, it distinguishes between synthetic and naturally occurring hallucinations, offering a cost-effective method for benchmarking detection models in production environments.

Table 7 summarizes major neuro-symbolic AI agent benchmarks and their key limitations.

11.2. Evaluation metrics

Recent surveys [152] identified the need for standardized verification and validation approaches in neuro-symbolic AI systems. Comprehensive evaluation requires metrics addressing multiple neuro-symbolic agent performance dimensions. *Task performance metrics* provide baseline assessments through success rate, efficiency measures, accuracy, and robustness. While essential, these metrics don't specifically address the quality of neuro-symbolic integration or distinguish between integrated performance and component contributions.

Integration quality metrics specifically evaluate neural-symbolic combination effectiveness through measures of neural-symbolic alignment (representation consistency), reasoning transparency (decision process comprehensibility), knowledge transfer between components (information flow efficiency), adaptation efficiency (representation modification ability), and semantic consistency (meaning maintenance across representations). These specialized metrics better capture neural-symbolic advantages but are less standardized and more challenging to implement consistently.

Agentic capability metrics assess system autonomy (independent operation ability), goal management (objective pursuit effectiveness), proactivity (problem anticipation), and adaptivity (novel situation adjustment). These higher-level capabilities often emerge from effective neuro-symbolic integration but require specialized evaluation approaches assessing qualities like persistence, flexibility, and goal-directedness across extended interactions.

Human-centric evaluation complements technical metrics by assessing human expectation alignment through explanation quality, trust measures, collaboration effectiveness, and value alignment. These evaluations provide crucial real-world utility insights but face standardization challenges due to the subjective nature of assessments and the influence.

Table 7
Neurosymbolic AI agent benchmarks and their limitations.

Benchmark	Focus	Key limitations
AgentBench [136]	Tool Use and Planning	Lacks multi-agent coordination, long-term memory evaluation
FOLIO [73]	Logical Reasoning	Limited scope versus human reasoning; prone to memorization
RSBench [149]	Concept Quality & Shortcuts	Focuses on arithmetic/logic; limited to specific architectures
LibreEval [150]	RAG Hallucination Detection	Primarily text-based; relies on LLM-as-a-Judge consensus
MDS-A [151]	Distribution Shift (Aerial)	Domain-specific (imagery); focuses on detection over reasoning
WildBench [137]	Real-world User Tasks	English-only; reproducibility issues; annotation inconsistencies
AgentClinic [94]	Multimodal Clinical Simulation	Poor simulation fidelity; subjective metrics; lacks longitudinal data
StrategyQA [138]	Multi-Step Reasoning	Focuses on answers, not reasoning validity
SOTOPIA [115]	Social Intelligence	Agents underperform humans in complex social scenarios

11.3. Proposed evaluation framework

Based on our analysis, we propose a comprehensive framework for evaluating neuro-symbolic AI agents integrating diverse metrics across multiple evaluation dimensions, and providing a structured approach for assessing neuro-symbolic agent performance and identifying areas.

This framework, shown in Fig. 8, incorporates three complementary evaluation dimensions: capability assessment (evaluating foundational abilities in perception, reasoning, learning, and action execution), integration evaluation (examining knowledge representation, reasoning processes, and learning mechanism integration), and human-centered assessment (evaluating explainability, trustworthiness, and collaborative effectiveness). The framework implements progression from controlled lab evaluations to real-world deployment assessments, providing a comprehensive picture of agent performance and limitations across contexts.

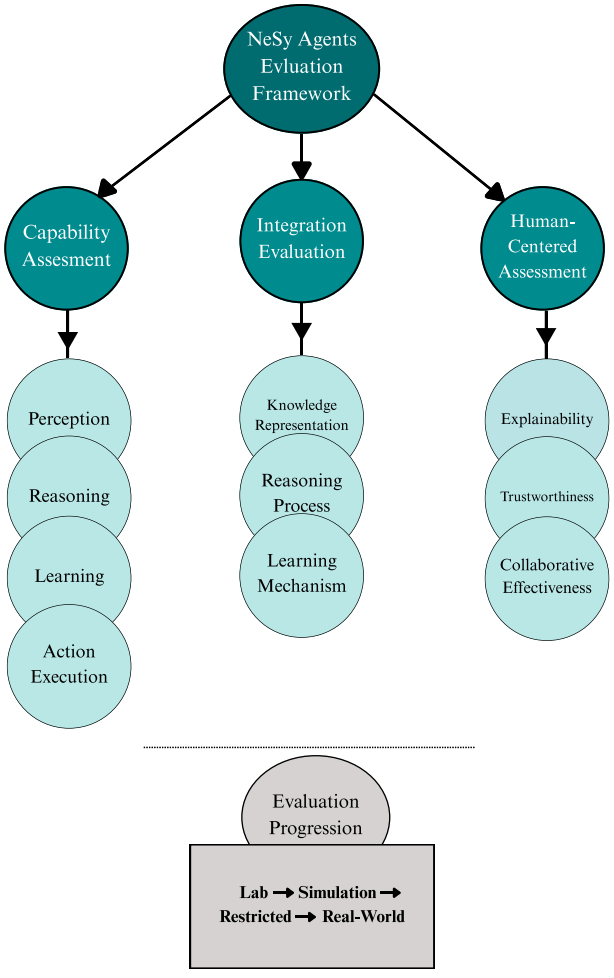

Fig. 8. Comprehensive evaluation framework for neuro-symbolic AI agents, assessing core capabilities, integration quality, and human-centered aspects.

TRAP-specific meta-cognitive evaluation metrics are provided in Appendix C, offering specialized assessment methods for systems adopting the TRAP framework [51].

11.4. Future benchmarking needs

Several critical benchmarking needs must be addressed in advancing neuro-symbolic agent assessment. The field requires significant advances in the following areas:

1. **Integration-Specific Benchmarks:** Standardized benchmarks specifically designed for evaluating neural-symbolic integration quality [126] should assess tasks requiring both pattern recognition and logical reasoning, information transfer between components, adaptability under knowledge changes, and reasoning transparency. These benchmarks must go beyond evaluating neural and symbolic capabilities separately to specifically target integration effectiveness.
2. **Standardized Metrics Development:** The field urgently needs standardized neuro-symbolic integration metrics, including quantitative neural-symbolic alignment measures, reasoning transparency metrics, metacognitive capability evaluation techniques, and comparative frameworks for assessing different integration approaches. Without consistent metrics, meaningful architectural approach comparisons remain challenging.
3. **Adaptability Evaluation Protocols:** Evaluation protocols for adaptability and transfer learning across diverse domains [153] are essential for understanding neuro-symbolic system performance beyond controlled settings. For instance, the MDS-A dataset [151] characterizes model performance under weather-induced distribution shifts in aerial imagery using Fréchet Inception Distance scores to quantify divergence. This approach enables systematic evaluation of test-time error detection and adaptation strategies, ensuring agents can identify failure points in safety-critical applications.
4. **Agentic Capability Metrics:** Comprehensive agentic capability metrics focusing on autonomy, goal-directedness, proactivity, and adaptability would significantly enhance evaluation effectiveness. These metrics should specifically address how well neuro-symbolic integration enhances agentic properties compared to purely neural or purely symbolic approaches.

Addressing these needs will enable more meaningful neuro-symbolic architecture comparisons and provide clearer future research guidance.

12. Case study: AlphaGeometry

AlphaGeometry [49] demonstrates how neuro-symbolic integration enhances reasoning capabilities beyond either approach independently through a neural-guided symbolic search architecture. This case study examines how AlphaGeometry implements key neuro-symbolic taxonomy components and broader implications for neuro-symbolic agent design.

12.1. Taxonomic analysis

AlphaGeometry implements a single-agent architecture with neural-guided symbolic search integration, combining neural language models' pattern recognition capabilities with symbolic theorem proving precision. The evolved AlphaGeometry2 architecture [154], illustrated in Fig. 9, extends this foundation with the SKEST framework for parallel search coordination.

From a taxonomy perspective, AlphaGeometry exemplifies several key integration patterns. The integration approach follows sequential integration (neural-to-symbolic) where the neural guide suggests promising deduction steps that the symbolic prover verifies for mathematical correctness. The neural-guided symbolic search approach follows a general algorithmic pattern enabling efficient exploration of large search spaces while maintaining formal correctness, as outlined in Algorithm 1.

In this approach, the neural component (N) provides heuristic guidance prioritizing promising search paths, while the symbolic component (S) ensures correctness through formal expansion rules. This integration creates a "best-of-both-worlds" scenario where neural pattern recognition dramatically improves search efficiency without sacrificing the formal guarantees that symbolic methods provide. The neural-guided symbolic search pattern represents one of the most successful integration approaches for complex reasoning tasks requiring both efficiency and correctness.

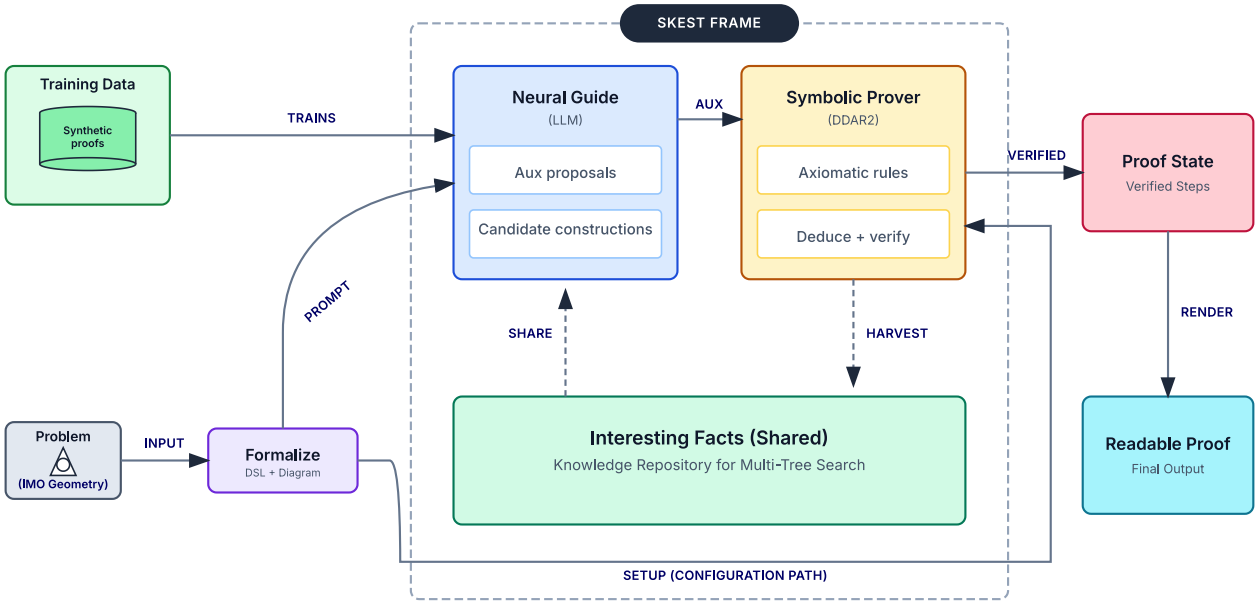

Fig. 9. AlphaGeometry2 architecture highlighting SKEST: parallel search trees share verified "interesting facts" through a shared knowledge repository, while a neural guide (LLM) proposes candidate constructions and the DDAR2 symbolic prover verifies correctness. The formalized problem specification feeds both components, enabling efficient multi-tree search with shared working memory.

Algorithm 1 Neural-guided symbolic search.

Require: Problem specification P, Neural guidance model N, Symbolic solver S
Ensure: Solution to problem P
Initialize search state s_0 based on P
Initialize fringe $F \leftarrow \{s_0\}$
while F not empty **do**
  Select state s from F with highest $N(s)$ (neural evaluation)
  Generate candidate next states $\{s'\}$ using S (symbolic expansion)
  for each candidate state s' **do**
    if s' is a solution **then**
      return s'
    end if
    Compute neural evaluation $N(s')$
    Add s' to F
  end for
end while
return No solution found

The system implements hybrid knowledge representation combining neural-encoded geometric patterns learned from synthetic examples with formal symbolic axioms defining Euclidean geometry rules. The learning mechanism uses self-supervised learning through synthetic proof generation, demonstrating how symbolic knowledge bootstraps neural learning even with limited human-created examples. The reasoning paradigm implements fast-slow thinking where the neural model (fast) generates candidate constructions based on pattern recognition while the symbolic prover (slow) verifies validity through formal proof.

Building on this foundational pattern, AlphaGeometry2 augments the neural-guided symbolic loop with SKEST (Shared Knowledge Enhanced Search Trees): multiple concurrent search trees exchange verified intermediate "interesting facts" through a shared knowledge repository (Fig. 9), effectively acting as a shared working memory across branches. Coupled with the faster DDAR2 prover, this orchestration improves sample efficiency, curbs redundant exploration, and boosts end-to-end geometry success on harder IMO instances.

12.2. Limitations and challenges

Despite impressive achievements, AlphaGeometry faces significant limitations. Domain-specific focus is limited to IMO-style Euclidean geometry problems, with authors acknowledging "scaling beyond Olympiad geometry remains an open question" [49]. This highlights the challenge of creating models that generalize beyond specific training domains, a recurring issue in neuro-symbolic systems implementing domain-specific integration approaches.

Training data generation required "nearly 1 billion" premises generated in parallel, with synthetic proof triples derived via symbolic traceback [49]. While enabling effective synthetic data learning, this approach requires significant computational resources and domain expertise, limiting applicability to domains where similar synthetic generation might be more challenging.

Reproducibility challenges further limit scientific and practical impact. While the official repository provides inference code, reproducing the complete training pipeline presents significant challenges for researchers with limited computational resources. System reliance on massive synthetic data generation complicates independent verification and extension.

Generalization concerns persist despite impressive benchmark performance. Despite solving 25/30 IMO problems (compared to 10/30 for pure symbolic approaches shown in Table 6), questions remain about the system ability to generalize to different types of mathematical reasoning or problem formats outside the training distribution. This limited generalization represents a common challenge for neuro-symbolic systems which excel in specific domains but struggle with capability transfer to new contexts.

12.3. Key insights for neurosymbolic design

AlphaGeometry provides valuable insights for effective neuro-symbolic agent design. Most significantly, it demonstrates how neural and symbolic components address each other's fundamental limitations—neural guidance improves search efficiency by focusing on promising directions [155] while symbolic verification ensures correctness and interpretability through formal proofs. This complementarity, mirroring architectural foundations laid by AlphaGo [129], creates a system leveraging each approach strengths while mitigating individual weaknesses.

System effectively preserves explicit domain knowledge (geometric axioms and theorems) while leveraging implicit patterns learned by neural components, maintaining both approaches' advantages. This integration enables both formal verification and intuitive pattern recognition, combining the reliability of symbolic methods with the flexibility of neural approaches—a synergy essential for robust complex agent reasoning [156].

Synthetic data generation approach demonstrates how systems generate their own training data through symbolic reasoning, enabling effective neural learning despite limited human-created examples. This self-supervised approach could generalize to other domains with formal structure but limited training data, potentially addressing the common challenge of applying neural methods to specialized domains with scarce labeled data. The efficacy of synthetic curricula is further validated by frameworks like AgentGen [106], which automate diverse planning environment generation enhancing agent capabilities without relying on human demonstrations.

From an agentic perspective, AlphaGeometry exhibits strong goal-directedness, employing neural guidance to efficiently explore vast search spaces while maintaining a focus. This demonstrates how neuro-symbolic integration enhances purposeful behavior in complex problem-solving domains by combining neural guidance efficiency with symbolic verification precision.

12.4. Future directions based on case study

Building on Section 6.5 identified research gaps, AlphaGeometry suggests several directions for advancing neuro-symbolic agents. Integrating explicit metacognitive capabilities would enhance the system ability to monitor and adjust reasoning strategies based on problem characteristics, addressing a significant meta-cognitive research gap. Enhancing explanation capabilities beyond formal proof steps would improve accessibility and utility for human users through higher-level conceptual insights explaining proof strategy intuition rather than just formal derivation steps. These enhanced explanations would make reasoning processes more accessible to non-experts and more useful for educational applications.

Extension to broader domains would demonstrate the generalizability of the neural-guided symbolic search approach [70]. By applying similar integration patterns to other formal domains like program synthesis, logical reasoning, or physics problem solving, researchers could assess how well this approach generalizes across different structured problem types. Improved computational efficiency would increase accessibility and deployment options by reducing the resources required for training and inference, enabling broader adoption in educational and research contexts where computational resources may be limited. The development of automated translation mechanisms between natural language problem statements and formal representations would enhance usability eliminating the need for manual translation.

Creating more autonomous variants identifying interesting problems rather than only addressing pre-specified challenges would enhance agentic capabilities. Such systems could actively explore problem

spaces discovering novel theorems or interesting conjectures rather than just proving given statements, addressing limited work connecting neuro-symbolic integration with agentic properties like autonomy and proactivity.

13. Open challenges

Despite significant advances in neuro-symbolic AI agents, several critical challenges persist that must be addressed to realize their full potential. This section identifies key open challenges in an enumerated fashion, providing a focused discussion of each limitation.

1. **Meta-Cognitive Capability Gap:** As revealed in Section 6.5 and Fig. 5, meta-cognition appears in only 5% of reviewed papers despite evidence from Section 6.4 that meta-cognitive capabilities contribute more to performance than sophisticated integration patterns alone. Current systems lack comprehensive self-assessment, strategy adaptation, and reflective reasoning mechanisms for autonomous operation in dynamic environments. The bootstrapping problem described in Section 10.4 compounds this issue, wherein LLMs' inherent self-assessment weaknesses create circular dependencies thereby limiting reliability.
2. **Explainability-Integration Disconnect:** While neuro-symbolic systems theoretically offer enhanced interpretability, only 28% of the surveyed literature addresses explainability as shown in Table 2. The intersection of explainability with knowledge representation appears in merely 4% of papers, indicating a severe gap in developing representations designed to enhance interpretability while maintaining neural flexibility. Recent reviews [14,43] confirm that most neuro-symbolic systems achieve only medium-low explainability despite the presence.
3. **Scalability and Computational Overhead:** Neurosymbolic systems face substantial computational requirements when combining large neural models with complex symbolic reasoning. As detailed in Section 5.1, advanced reasoning paradigms like Tree of Thoughts require 5–100× more tokens than Chain-of-Thought [89], while training costs reach $78M for GPT-4 and $191M for Gemini Ultra [130]. Knowledge integration complexity escalates significantly as knowledge bases expand, with few approaches addressing the combinatorial explosion in reasoning paths [127].
4. **Integration Methodology Standardization:** The absence of standardized methodologies for combining neural and symbolic components hampers progress [125]. As shown in Tables 4 and 5, different integration patterns exhibit varying strengths across dimensions, yet few papers provide metrics assessing semantic consistency between neural and symbolic representations or formal verification of integration quality. This methodological gap prevents systematic comparison and principled cross-domain integration design.
5. **Benchmark Inadequacy and Evaluation Gaps:** Current benchmarks inadequately assess neuro-symbolic integration quality, as documented in Table 7. Most evaluations focus on either neural or symbolic capabilities in isolation rather than their synergy [126]. AgentBench lacks multi-agent coordination assessment [136], FOLIO exhibits memorization tendencies [73], and AgentClinic demonstrates poor simulation fidelity [94]. Metrics for integration quality, cross-component knowledge transfer, and adaptive capabilities remain underdeveloped.
6. **Symbol Grounding Persistence:** Despite modular design advances [35], connecting abstract symbolic representations to real-world perceptions remains challenging, particularly for novel objects in embodied contexts. As illustrated in Fig. 6, bridging continuous perceptual inputs and discrete symbolic representations creates fundamental difficulties [86,139]. Most robotic systems are evaluated in controlled environments with limited variability, raising questions about real-world robustness.
7. **Domain Generalization Limitations:** Systems excelling in specific domains struggle with transferring capabilities to new contexts. AlphaGeometry's authors acknowledge "scaling beyond Olympiad geometry remains an open question" [49], while Agent Q struggles with novel web interfaces unseen during training [16]. Even LINC shows brittleness with disordered premises and distractors [73], aligning with broader transformer compositionality limits creating "capability cliffs" [102].
8. **Reproducibility Barriers:** Significant computational requirements create reproducibility barriers limiting independent verification. AlphaGeometry required nearly 1 billion synthetically generated premises [49], MetaGPT demands manual SOP design for new tasks [50], and Reflexion faces high API costs [42]. These barriers hinder progress by preventing researchers from effectively building on prior work.
9. **Ethical Emergence and Bias Propagation:** Integration complexity creates unpredictable interactions raising ethical concerns. Neural components may learn data biases while symbolic rules encode assumptions [78], creating dynamics that are difficult to analyze. Systems like CulturePark and SOTOPIA show emergent behaviors risking unintended biases [114,115]. Privacy vulnerabilities are heightened as symbolic components make sensitive information explicit [43].
10. **Agentic Property Formalization Gap:** Insufficient work connects formal agentic properties (autonomy, goal-directedness, proactivity) with neuro-symbolic integration mechanisms. While frameworks like Agent Q [16], GoalAct [48], and BNE-Q [118] demonstrate potential, systematic research characterizing how integration patterns enhance specific agentic capabilities remains sparse, limiting the principled design of autonomous systems.
11. **Human-Agent Collaboration Complexity:** Neurosymbolic systems present unique collaboration challenges despite transparency advantages. The "automation surprise" problem persists where users become confused by unexpected agent behavior. As discussed in Section 8.4, balancing predictability with adaptability requires careful integration pattern selection, yet principled guidelines for choosing patterns based on collaboration requirements remain underdeveloped.
12. **Long-Horizon Task Brittleness:** Systems struggle to maintain coherence across extended interactions despite advances such as GITM's memory architecture [17] and narrative memory integration [39,40]. Reflexion's sliding window approach limits the ability to leverage temporally distant insights [105], while StarPO addresses but doesn't fully resolve "Echo Trap" phenomena where agents enter self-repeating loops [107].

These enumerated challenges represent fundamental obstacles requiring a concerted interdisciplinary effort combining insights from cognitive science, formal methods, machine learning, and human-computer interaction to advance neuro-symbolic agent capabilities toward robust, trustworthy, autonomous systems.

14. Future research directions

Building on open challenges identified in Section 13, we propose promising directions for future research that move beyond addressing current limitations to envision next-generation neuro-symbolic systems with enhanced capabilities.

14.1. Enhanced integration methodologies

Fundamental neuro-symbolic integration challenges bridge the gap between neural flexibility and symbolic precision while maintaining gradient-based optimization. Advances in differentiable programming

offer promising solutions through differentiable symbolic operation implementations [67,157] enabling gradient flow through previously discrete operations. These implementations approximate discrete operations with continuous alternatives supporting backpropagation while preserving symbolic semantics, as formalized in Appendix A.

Neural architectures specifically designed to complement symbolic reasoning could provide another approach to implementing structures naturally aligning with symbolic representations, including attention mechanisms focusing on symbolic elements or relational inductive biases capturing symbolic structure. Gradient-based optimization of integrated neural-symbolic systems would enable end-to-end training of complete neuro-symbolic pipelines, potentially improving both performance and coherence. This approach would allow neural and symbolic components to co-adapt during training, creating more effective integration than separate individual component optimization.

Future research should explore adaptive integration selection mechanisms that dynamically choose optimal integration patterns based on task characteristics, computational constraints, and performance requirements. As shown in Table 5, different patterns excel across different dimensions—developing meta-learning approaches that automatically select and configure integration strategies would significantly enhance system versatility.

14.2. Advanced meta-cognitive frameworks

Addressing a significant meta-cognitive research gap (5% of papers) represents a major opportunity for advancing neuro-symbolic agent capabilities [14]. Future research should develop integrated architectures implementing comprehensive meta-cognitive capabilities across all TRAP dimensions illustrated in Fig. 7: self-monitoring mechanisms evaluating reasoning quality, strategy selection and adaptation based on task characteristics, resource allocation optimization across components, and knowledge integration management ensuring consistent representation updates [51].

Transfer learning approaches for meta-cognitive capabilities would enhance system flexibility enabling cross-domain generalization. For domain transfer $D_{\text{source}} \rightarrow D_{\text{target}}$, TRAP components can be adapted through component-specific transfer functions encoding domain adaptation knowledge. This formulation enables: *Abstract Strategy Transfer* where meta-cognitive strategies learned in one domain transfer to others by abstracting task-independent monitoring patterns; *Compositional Meta-Cognition* where complex capabilities emerge through the composition of transferred components; *Continual Meta-Cognitive Improvement* where architectures supporting continual learning update components incrementally without catastrophic forgetting.

Developing external grounding mechanisms to address the bootstrapping problem identified in Section 10.4 represents a critical direction. Future systems should incorporate independent verification sources—symbolic knowledge base checks, environment feedback, multimodal sensor redundancy—breaking circular self-assessment dependencies while maintaining TRAP's structured monitoring capabilities.

14.3. Hierarchical agentic architectures

Building on hierarchical planning advances demonstrated in GoalAct [48] and HLA [103], future research should explore more sophisticated hierarchical architectures combining symbolic task decomposition with neural execution. Fig. 10 illustrates a potential integrated architecture addressing the fundamental tension between planning coherence and execution flexibility.

This architecture combines autonomy mechanisms from Agent Q for neural-guided symbolic search, hierarchical goal management from GoalAct maintaining coherent objectives, structured memory from GITM for efficient plan reuse, and coordination protocols from AutoGen/MetaGPT for multi-agent collaboration. Bidirectional flows between components enable continuous adaptation and feedback, creating a system that both maintains long-term goals and responds to environmental changes.

Developing general-purpose symbolic decomposition frameworks would enable systems to generate appropriate task hierarchies across diverse domains without requiring domain-specific decomposition strategies. These frameworks would identify natural task hierarchical structures and create appropriate goal decompositions based on task characteristics, enhancing cross-domain applicability and reducing domain-specific engineering needs. Designing feedback loops between execution and planning would enable dynamic goal adjustment based on observed outcomes. Integrating formal goal verification mechanisms would ensure that subgoal achievement genuinely contributes to overarching objectives.

14.4. Standardized evaluation frameworks

Developing robust evaluation frameworks specifically designed for neuro-symbolic agents is essential for systematic progress assessment [126]. Building on our proposed framework in Fig. 8, future research

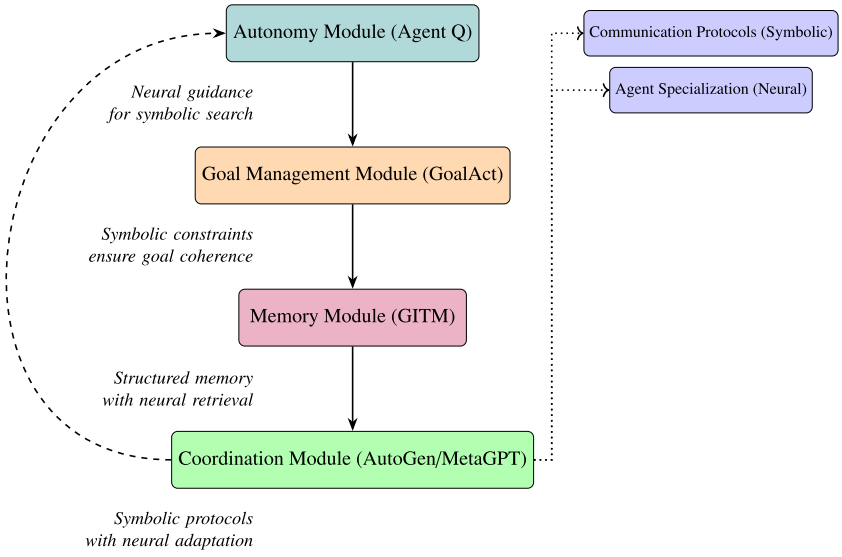

Fig. 10. Proposed integrated neuro-symbolic agentic architecture combining autonomy (Agent Q), goal management (GoalAct), structured memory (GITM), and coordination (AutoGen/MetaGPT) with bidirectional information flows balancing long-term goal coherence with adaptive execution.

should establish: *Integration-focused benchmarks* specifically assessing neural-symbolic integration quality through tasks requiring both pattern recognition and logical reasoning; *Standardized integration metrics* providing consistent evaluation including quantitative neural-symbolic alignment measures, reasoning transparency metrics, and metacognitive capability evaluation techniques; *Cross-domain transfer protocols* evaluating generalization beyond training distributions; *Agentic capability assessments* measuring autonomy, goal-directedness, proactivity, and adaptability enhancements from neuro-symbolic integration.

Community-driven benchmark development initiatives could establish shared evaluation infrastructure similar to successful efforts in other AI domains. These initiatives should prioritize: diverse task coverage spanning reasoning, perception, planning, and social interaction; multilingual and multicultural evaluation addressing biases in current English-centric benchmarks like WildBench [137]; longitudinal assessment protocols evaluating performance over extended interactions critical for healthcare applications [94]; open-source implementations facilitating reproducibility and community participation.

14.5. Multimodal and embodied neuro-symbolic agents

Extending neuro-symbolic integration to multimodal agents handling diverse input modalities (vision, audio, touch, proprioception) represents a frontier for creating truly general-purpose systems. As discussed in Section 5.10, current Large Multimodal Agents [7] demonstrate potential but face scalability and grounding challenges.

Future research should explore: *Cross-modal symbol grounding* mechanisms for learning consistent symbolic representations across multiple sensory modalities, addressing the symbol grounding problem illustrated in Fig. 6 through redundant multimodal evidence; *Embodied common sense reasoning* combining neural perception with symbolic world models enabling physical reasoning about object interactions, stability, and causality; *Situated learning architectures* where agents learn symbol meanings through environmental interaction following approaches like Voyager [141], enabling robust generalization to novel situations.

Developing neuro-symbolic frameworks for human-robot collaboration would enhance safety and transparency in physical interaction scenarios. These frameworks could combine neural gesture recognition with symbolic task representations, enabling robots to interpret ambiguous human communication while maintaining an explicit shared understanding of collaborative goals.

14.6. Ethical and responsible development

As neuro-symbolic agents grow more capable, ethical considerations [158] become increasingly important ensuring responsible development and deployment [8]. Future research should address: *Fairness in integrated systems* through techniques detecting and mitigating bias in neural-symbolic interactions, methods incorporating ethical constraints in symbolic components, and approaches verifying fairness properties across integrated systems; *Privacy-preserving neuro-symbolic learning* enabling learning from sensitive data without exposing private information through differential privacy or federated learning adapted to neuro-symbolic systems; *Controlled information flow* preventing sensitive information leakage across neural-symbolic boundaries while maintaining integration benefits.

Developing formal verification methods for neuro-symbolic systems would enable proving safety and ethical properties before deployment. These methods could leverage symbolic components' explicit structure for tractable verification while accounting for neural components' learned behaviors through abstraction and approximation techniques. Highly autonomous agent risks warrant careful consideration, particularly as neuro-symbolic approaches enhance autonomy. Semi-autonomous designs with human oversight may provide a safer approach balancing efficiency with safety [18].

14.7. Neuro-symbolic foundation models

Current foundation models lack explicit symbolic reasoning capabilities despite their impressive performance. Developing neuro-symbolic foundation models that combine large-scale neural pretraining with structured symbolic knowledge represents an ambitious long-term direction. These models could: integrate knowledge graphs at the foundation model scale providing grounding for factual reasoning; implement differentiable symbolic operations as primitive components enabling complex reasoning composition; and support modular skill acquisition where new symbolic reasoning capabilities can be added without requiring full retraining.

Such foundation models would address several current limitations simultaneously: improved factual consistency through symbolic knowledge grounding, enhanced reasoning transparency via explicit symbolic traces, better sample efficiency in specialized domains through symbolic prior knowledge, and stronger compositional generalization through symbolic operation combination. While technically challenging, the convergence of advances in differentiable programming, knowledge representation, and large-scale training makes this direction increasingly feasible.

14.8. Cross-domain knowledge transfer

Enabling effective knowledge transfer across domains represents a key challenge limiting current systems. Future research should develop: *Abstract symbolic primitives* identifying domain-independent reasoning patterns transferable across applications; *Meta-learning frameworks* for neuro-symbolic integration allowing systems to learn how to integrate neural and symbolic components effectively from experience across multiple domains; *Compositional knowledge transfer* mechanisms enabling the combination of knowledge from multiple source domains for new target domains.

AlphaGeometry's synthetic data generation approach [49] suggests a promising direction—developing general frameworks for synthetic curriculum generation across domains could dramatically reduce data requirements for neuro-symbolic systems. As demonstrated by AgentGen [106], automated environment and trajectory generation enable agents to acquire robust capabilities without extensive human demonstration.

15. Conclusion

This comprehensive survey establishes neuro-symbolic AI agents as a transformative paradigm synergizing neural networks' pattern recognition with symbolic reasoning's logical structure. Our systematic PRISMA-based analysis of 178 papers indicates that neuro-symbolic integration consistently outperforms pure approaches across diverse domains—achieving 23% improvements in robotic task completion, 95.4% autonomous navigation success versus 18.6% neural baselines, and order-of-magnitude reductions in sample complexity. Through a comprehensive taxonomy spanning architectural configurations and integration dimensions, we demonstrate how different patterns—sequential, parallel, end-to-end differentiable, unified representation—exhibit complementary strengths addressable to specific application requirements. Performance comparisons in Table 6 substantiate that advantages arise from fundamental complementarity: neural components address symbolic brittleness in uncertain environments while symbolic components enhance neural interpretability and systematic reasoning.

However, critical challenges persist demanding immediate research attention. Meta-cognitive capabilities, appearing in only 5% of the literature despite demonstrating greater performance impact than sophisticated integration patterns alone, represent a severely underexplored frontier. The explainability-integration disconnect—with merely 4% of papers addressing both knowledge representation and explainability—prevents the realization of neuro-symbolic systems' theoretical transparency advantages. Scalability constraints, integration methodology standardization gaps, benchmark inadequacies enumerated in

Section 13, and persistent symbol grounding difficulties constitute fundamental obstacles. Our proposed future directions—advanced metacognitive frameworks via TRAP-inspired architectures, hierarchical agentic systems illustrated in Fig. 10, standardized evaluation protocols from Fig. 8, and neuro-symbolic foundation models—provide a structured roadmap addressing these challenges. By combining insights from cognitive science, formal methods, and human-computer interaction, the community can advance toward truly autonomous, transparent, and trustworthy intelligent systems collaborating effectively with humans across complex real-world applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval

This article does not contain any studies with human participants or animals.

Appendix A. Mathematical foundations of neuro-symbolic integration

This appendix provides mathematical formulations for key neuro-symbolic integration patterns discussed in Section 3.3, supporting the conceptual overview with formal definitions.

A.1. Sequential integration

Sequential integration connects neural and symbolic components in a feed-forward pipeline. In the neural-to-symbolic direction:

$$\begin{aligned} Z_{\text{neural}} &= f_{\text{neural}}(X) \\ Y &= f_{\text{symbolic}}(Z_{\text{neural}}) \end{aligned} \tag{A.1}$$

where X represents input data, Z_{neural} is an intermediate neural representation, and Y is the output after symbolic processing. This pattern commonly appears in frameworks like Agent Q [16] where neural perception feeds symbolic planning.

Reverse sequential pattern (symbolic-to-neural) follows:

$$\begin{aligned} Z_{\text{symbolic}} &= f_{\text{symbolic}}(X) \\ Y &= f_{\text{neural}}(Z_{\text{symbolic}}) \end{aligned} \tag{A.2}$$

This enables symbolic constraints to guide neural generation, ensuring that outputs satisfy logical requirements.

A.2. Parallel integration

Parallel integration processes input through both neural and symbolic pathways simultaneously, combining their outputs:

$$\begin{aligned} Z_{\text{neural}} &= f_{\text{neural}}(X) \\ Z_{\text{symbolic}} &= f_{\text{symbolic}}(X) \\ Y &= g(Z_{\text{neural}}, Z_{\text{symbolic}}) \end{aligned} \tag{A.3}$$

where g is a fusion function combining outputs from both streams. Common fusion functions include:

- *Weighted averaging*: $g(Z_{\text{neural}}, Z_{\text{symbolic}}) = \alpha Z_{\text{neural}} + (1-\alpha) Z_{\text{symbolic}}$
- *Logical filtering*: $g(Z_{\text{neural}}, Z_{\text{symbolic}}) = Z_{\text{neural}} \odot \mathbb{I}[Z_{\text{symbolic}}]$
- *Learned combination*: $g(Z_{\text{neural}}, Z_{\text{symbolic}}) = h_\theta([Z_{\text{neural}}, Z_{\text{symbolic}}])$

Hierarchical Language Agents [103] employ this pattern combining fast neural responses with slower symbolic reasoning.

A.3. End-to-end differentiable architectures

End-to-end differentiable architectures embed symbolic operations within neural networks enabling gradient-based optimization:

$$\begin{aligned} Y &= f_{\text{neuro-symbolic}}(X) \\ &= f^{(n)}_{\text{neural}} \circ f^{(n-1)}_{\text{diff-symbolic}} \circ \ldots \circ f^{(1)}_{\text{neural}}(X) \end{aligned} \tag{A.4}$$

where $f_{\text{diff-symbolic}}$ represents differentiable symbolic operation implementations. These typically use continuous relaxations of discrete logical operations [77]:

$$\begin{aligned} \text{AND}_\tau(a,b) &= \sigma\left(\frac{a+b-1}{\tau}\right) \\ \text{OR}_\tau(a,b) &= \sigma\left(\frac{a+b}{\tau}\right) \\ \text{NOT}_\tau(a) &= 1-a \end{aligned} \tag{A.5}$$

where σ is the sigmoid function and τ is the temperature parameter controlling approximation sharpness. As $\tau \to 0$, these operations approach discrete logical functions.

For gradient computation through symbolic layers:

$$\frac{\partial \mathcal{L}}{\partial \theta_{\text{neural}}} = \frac{\partial \mathcal{L}}{\partial Y} \cdot \frac{\partial Y}{\partial f_{\text{diff-symbolic}}} \cdot \frac{\partial f_{\text{diff-symbolic}}}{\partial \theta_{\text{neural}}} \tag{A.6}$$

where $\mathcal{L}$ represents the task loss and θ_{neural} denotes the neural network parameters.

A.4. Unified representation approaches

Unified representation approaches like ZeroC [47] establish one-to-one mappings between symbolic structures and neural energy functions [62]. For concept c represented as symbolic graph $G_c = (V_c, E_c)$ with nodes V_c (base concepts) and edges E_c (relations):

$$E(c,x) = \sum_{v_i \in V_c} E_{\text{base}}(v_i, x) + \sum_{e_j \in E_c} E_{\text{rel}}(e_j, x) \tag{A.7}$$

where $E(c,x)$ is the energy function for concept c applied to input x, composed of base concept energies $E_{\text{base}}(v_i, x)$ and relation energies $E_{\text{rel}}(e_j, x)$.

Recognition probability for input x belonging to concept c:

$$p(c|x) = \frac{\exp(-E(c,x))}{\sum_{c' \in C} \exp(-E(c',x))} \tag{A.8}$$

where C represents set of all concepts. This formulation enables zero-shot concept recognition by composing existing concepts according to symbolic descriptions:

$$c_{\text{new}} = \text{Compose}(c_1, c_2, \ldots, c_k; R) \tag{A.9}$$

where R specifies the relational structure combining base concepts $\{c_1, \ldots, c_k\}$.

A.5. Neural-guided symbolic search

Neural-guided symbolic search, exemplified by AlphaGeometry and detailed in Algorithm 1, combines neural heuristic evaluation with symbolic state expansion. The search process maintains a priority queue F of states ranked by neural evaluation:

$$s^* = \arg,\ \max_{s \in F} N(s) \tag{A.10}$$

where $N(s)$ is neural network predicting state s quality. For each selected state, symbolic expansion generates successor states:

$$\text{Successors}(s) = \{s' | s \xrightarrow{a} s', a \in \mathcal{A}(s)\} \tag{A.11}$$

where $\mathcal{A}(s)$ represents set of valid symbolic actions from state s. This integration achieves efficiency gain:

$$\text{Efficiency} = \frac{|\text{States}_{\text{exhaustive}}|}{|\text{States}_{\text{neural-guided}}|} \tag{A.12}$$

measuring the reduction in explored states compared to exhaustive symbolic search.

Appendix B. TRAP framework mathematical formulations

This appendix provides detailed mathematical formulations for the TRAP framework components referenced in Section 10.2, extending the conceptual definitions with formal specifications.

B.1. TRAP component functions

The TRAP framework characterizes four meta-cognitive targets through specific placements of meta-cognitive function g relative to the base model f_θ:

B.1.1. Transparency function

Transparency-oriented assessment generates explanations based on model outputs and parameters:

$$g_T(f_\theta(x), \theta) = \text{Explain}(f_\theta(x), \theta, \mathcal{K}_{\text{domain}}) \tag{B.1}$$

where $\mathcal{K}_{\text{domain}}$ represents domain knowledge used for explanation generation. The transparency score evaluates explanation quality:

$$\text{Score}_T = \mathbb{E}_{x\sim\mathcal{D}}\left[\text{Fidelity}(g_T) \times \text{Comprehend}(g_T) \times \text{Complete}(g_T)\right] \tag{B.2}$$

where fidelity measures explanation accuracy, comprehensibility measures human understanding, and completeness measures information coverage.

B.1.2. Reasoning function

Reasoning quality assessment evaluates logical consistency and inference validity:

$$g_R(\theta, \mathcal{H}) = \text{Assess}(\text{Inferences}(\theta), \mathcal{H}) \tag{B.3}$$

where $\mathcal{H}$ represents performance history and $\text{Inferences}(\theta)$ extracts reasoning steps from model. The reasoning score:

$$\text{Score}_R = \frac{|\text{ValidInferences}|}{|\text{TotalInferences}|} \times (1 - \text{ContradictionRate}) \tag{B.4}$$

B.1.3. Adaptation function

Adaptation assessment identifies error patterns and strategy effectiveness:

$$g_A(f_\theta(x), \theta, \mathcal{E}) = \text{Diagnose}(\mathcal{E}, \theta) \oplus \text{Prescribe}(\mathcal{E}, \theta) \tag{B.5}$$

where $\mathcal{E}$ represents detected errors, Diagnose identifies adaptation needs, and Prescribe suggests modifications. Adaptation leads to updated model:

$$f'_{\theta'}(x) = \text{Apply}(f_\theta(x), g_A(f_\theta(x), \theta, \mathcal{E})) \tag{B.6}$$

with parameters:

$$\theta' = \theta + \eta \cdot \Delta\theta_{g_A} \tag{B.7}$$

B.1.4. Perception function

Perception confidence assessment evaluates input quality and grounding reliability:

$$g_P(x, C) = \text{Confidence}(x, C) \times \text{Grounding}(x, \mathcal{K}_{\text{world}}) \tag{B.8}$$

where C represents context and $\mathcal{K}_{\text{world}}$ represents world knowledge. Low perception confidence triggers attention mechanisms:

$$\alpha_{\text{attend}} = \begin{cases} 1 & \text{if } g_P(x, C) > \tau_P \\ \text{sigmoid}(w \cdot g_P(x, C)) & \text{otherwise} \end{cases} \tag{B.9}$$

B.2. Unified TRAP meta-cognitive state

Following TRAP's four-way decomposition, we aggregate component monitors into a unified meta-cognitive state:

$$M_{\text{TRAP}}(s, o, f_O) = \begin{bmatrix} g_T(f_O(s), \theta) \\ g_R(\theta, \mathcal{H}) \\ g_A(f_O(s), \theta, \mathcal{E}) \\ g_P(o, C) \end{bmatrix} \tag{B.10}$$

where s is the state, o is the observation, $\mathcal{H}$ is the performance history, $\mathcal{E}$ is the detected errors, and C is the context. Each component contributes distinct meta-cognitive signals consistent with TRAP's schematic placements.

B.3. TRAP-inspired adaptive control

TRAP signals guide integration pattern selection through policy:

$$\pi_{\text{TRAP}}(s_t) = \arg,\max_{p\in\mathcal{P}} \sum_{i\in\{T,R,A,P\}} w_i \cdot \text{Utility}_i(p, M_{\text{TRAP}}(s_t)) \tag{B.11}$$

where $\mathcal{P} = \{\text{Sequential, Parallel, E2E, Unified}\}$ denotes available patterns. Component utilities encode task-dependent preferences:

$$\text{Utility}_T(p, M) = \mathbb{I}[\text{HasTrace}(p)] \cdot g_T(M) \tag{B.12}$$

$$\text{Utility}_R(p, M) = \text{LogicalConsist}(p) \cdot g_R(M) \tag{B.13}$$

$$\text{Utility}_A(p, M) = \text{Flexibility}(p) \cdot g_A(M) \tag{B.14}$$

$$\text{Utility}_P(p, M) = \text{GroundQuality}(p) \cdot g_P(M) \tag{B.15}$$

B.4. External grounding for TRAP components

To mitigate circular self-evaluation identified in Section 10.4, external grounding channels blend internal and external verification:

$$g_i^{\text{ext}}(x, f) = \alpha_i \cdot g_i^{\text{internal}}(x, f) + (1 - \alpha_i) \cdot V_i^{\text{external}}(x) \tag{B.16}$$

where external verifiers provide independent validation:

$$V_T^{\text{ext}}(x) = \text{RuleCheck}(\mathcal{K}_{\text{symbolic}}, g_T(f(x))) \tag{B.17}$$

$$V_R^{\text{ext}}(x) = \text{LogicVerify}(\mathcal{L}_{\text{formal}}, g_R(\theta)) \tag{B.18}$$

$$V_A^{\text{ext}}(x) = \text{EnvFeedback}(\text{env}, f(x)) \tag{B.19}$$

$$V_P^{\text{ext}}(x) = \text{SensorRedundancy}(\text{multimodal}, x) \tag{B.20}$$

Confidence weight α_i adapts based on historical agreement:

$$\alpha_i(t+1) = \alpha_i(t) + \beta \cdot \text{Agreement}(g_i^{\text{internal}}, V_i^{\text{external}}) \tag{B.21}$$

B.5. Meta-cognitive gradient correction

TRAP adaptation signals modulate learning when performance degrades:

$$\Delta\theta_t = \alpha \cdot \nabla_\theta \mathcal{L}_{\text{task}}(x_t, y_t) + \beta \cdot \nabla_\theta \mathcal{L}_{\text{meta}}(M_{\text{TRAP}}, \theta_t) \tag{B.22}$$

where meta-cognitive loss penalizes undesirable states:

$$\mathcal{L}_{\text{meta}}(M, \theta) = \sum_{i\in\{T,R,A,P\}} \lambda_i \|g_i(M) - g_i^{\text{target}}\|^2 + \gamma \cdot \mathcal{L}_{\text{coherence}}(M) \tag{B.23}$$

with coherence loss ensuring consistent meta-cognitive signals:

$$\mathcal{L}_{\text{coherence}}(M) = \sum_{i\neq j} \|\text{Correlation}(g_i, g_j) - C_{ij}^{\text{expected}}\|^2 \tag{B.24}$$

Appendix C. TRAP-specific meta-cognitive evaluation

This appendix provides specialized evaluation metrics for systems adopting the TRAP framework [51], referenced from Section 11.3. Each TRAP dimension requires tailored assessment aligned with its functional role.

C.1. Transparency evaluation

Transparency score evaluates explanation quality across multiple dimensions:

$$\text{Transparency}_{\text{score}} = \frac{1}{N}\sum_{i=1}^{N}\text{Fidelity}(g_T(f(x_i))) \cdot \text{Comprehend}(g_T(f(x_i))) \tag{C.1}$$

where:

- *Fidelity* measures explanation accuracy relative to the actual decision process, computed as the alignment between explanation steps and the execution trace
- *Comprehensibility* measures human understanding through user studies or proxy metrics like explanation length and complexity

Detailed transparency assessment includes:

$$\text{Fidelity}(e) = \frac{|\text{ExplanSteps}(e) \cap \text{ActualSteps}(f)|}{|\text{ActualSteps}(f)|} \tag{C.2}$$

$$\text{Comprehend}(e) = \exp\left(-\frac{\text{Length}(e)}{\tau_{\text{length}}}\right) \cdot \text{Readability}(e) \tag{C.3}$$

C.2. Reasoning evaluation

Reasoning score assesses logical consistency and inference validity:

$$\text{Reasoning}_{\text{score}} = \frac{|\text{CorrectReflections}|}{|\text{TotalReflections}|} \cdot \text{ConsistencyScore} \tag{C.4}$$

where consistency score measures logical coherence:

$$\text{ConsistencyScore} = 1 - \frac{|\text{Contradictions}|}{|\text{InferencePairs}|} \tag{C.5}$$

Inference pair (i_1, i_2) is considered contradictory if:

$$\exists i_1, i_2 : (\text{Conclusion}(i_1) \equiv \neg\text{Conclusion}(i_2)) \wedge (\text{Context}(i_1) = \text{Context}(i_2)) \tag{C.6}$$

C.3. Adaptation evaluation

Adaptation score quantifies improvement from meta-cognitive adjustment:

$$\text{Adaptation}_{\text{score}} = \frac{\text{Perf}_{\text{post}} - \text{Perf}_{\text{pre}}}{\max(\text{Perf}_{\text{baseline}}, \epsilon)} \cdot (1 - \text{AdaptLatency}) \tag{C.7}$$

where:

- $\text{Perf}_{\text{post}}$ measures performance after adaptation
- Perf_{pre} measures performance before adaptation
- $\text{Perf}_{\text{baseline}}$ provides reference performance
- $\text{AdaptLatency} = \frac{\text{Time}_{\text{adapt}}}{\text{Time}_{\text{total}}}$ penalizes slow adaptation

Adaptation efficiency across multiple episodes:

$$\text{AdaptEfficiency} = \frac{1}{K}\sum_{k=1}^{K}\frac{\Delta\text{Perf}_k}{\text{Cost}_{\text{adapt},k}} \tag{C.8}$$

C.4. Perception evaluation

Perception score assesses input quality and grounding reliability:

$$\text{Perception}_{\text{score}} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{|\text{Misground}(x_i)|}{|\text{Total}(x_i)|}\right) \cdot \mathbb{I}[g_P(x_i) > \tau_{\text{conf}}] \tag{C.9}$$

where:

- $\text{Misground}(x_i)$ counts symbols incorrectly grounded in the input x_i
- $\mathbb{I}[g_P(x_i) > \tau_{\text{conf}}]$ indicates satisfaction of the confidence threshold

Grounding accuracy measured by semantic consistency:

$$\text{GroundAccuracy} = \frac{\sum_{s\in S}\text{Similarity}(\text{Percept}(s), \text{GroundTruth}(s))}{|S|} \tag{C.10}$$

C.5. Aggregate TRAP effectiveness

Overall TRAP effectiveness uses the geometric mean encouraging balanced component performance:

$$\text{TRAP}_{\text{overall}} = \sqrt[4]{\text{Trans}_{\text{score}} \cdot \text{Reason}_{\text{score}} \cdot \text{Adapt}_{\text{score}} \cdot \text{Percept}_{\text{score}}} \tag{C.11}$$

This formulation penalizes systems excelling in only a subset of dimensions while rewarding balanced capabilities. An alternative weighted formulation allows task-specific emphasis:

$$\text{TRAP}_{\text{weighted}} = \prod_{i\in\{T,R,A,P\}}(\text{Score}_i)^{w_i} \tag{C.12}$$

where $\sum_i w_i = 1$ and weights are determined by task requirements.

C.6. Comparative TRAP assessment

For comparing systems with different TRAP implementations, the normalized improvement metric:

$$\text{TRAP}_{\text{improve}} = \frac{\text{TRAP}^{\text{system}}_{\text{overall}} - \text{TRAP}^{\text{baseline}}_{\text{overall}}}{\text{TRAP}^{\text{optimal}}_{\text{overall}} - \text{TRAP}^{\text{baseline}}_{\text{overall}}} \tag{C.13}$$

where baseline represents a system without meta-cognitive capabilities and optimal represents the theoretical maximum achievable TRAP scores.

C.7. Temporal TRAP dynamics

For evaluating the TRAP component evolution during learning:

$$\text{TRAP}_{\text{trajectory}}(t) = \begin{bmatrix}\text{Score}_T(t)\\ \text{Score}_R(t)\\ \text{Score}_A(t)\\ \text{Score}_P(t)\end{bmatrix} \tag{C.14}$$

Convergence rate measures meta-cognitive development speed:

$$\text{ConvergenceRate} = \min_i \frac{1}{t_i^*} \tag{C.15}$$

where $t_i^* = \min\{t : \text{Score}_i(t) > \tau_i\}$ indicates the time when component i reaches the threshold τ_i.

C.8. Meta-cognitive transfer assessment

For domain transfer $D_{\text{source}} \to D_{\text{target}}$, transfer efficiency:

$$\text{TransferEff} = \frac{\text{TRAP}^{D_{\text{target}}}_{\text{overall}}(\text{with transfer})}{\text{TRAP}^{D_{\text{target}}}_{\text{overall}}(\text{from scratch})} \tag{C.16}$$

Component-specific transfer success:

$$\text{TransferSuccess}_i = \frac{\text{Score}_i^{D_{\text{target}}}(\text{initial})}{\text{Score}_i^{D_{\text{source}}}(\text{final})} \tag{C.17}$$

These metrics quantify how effectively meta-cognitive capabilities learned in the source domain transfer to the target domain, which is essential for evaluating generalization as discussed in Section 14.2.

Data availability

No data was used for the research described in the article.

References

[1] H.A. Kautz, The third AI summer: AAAI Robert S. Engelmore memorial lecture, AI Magazine 43 (1) (2022) 105–125.

[2] A.D. Garcez, L.C. Lamb (Eds), Neuro-Symbolic Artificial Intelligence: the State of the Art, Frontiers in Artificial Intelligence and Applications, vol. 363, IOS Press, 2023.

[3] S. Bader, P. Hitzler, Dimensions of neural-symbolic integration-a structured survey, arXiv preprint arXiv:cs/0511042 2005.

[4] D.B. Acharya, K. Kuppan, B. Divya, Agentic AI: autonomous intelligence for complex goals–a comprehensive survey, IEEE Access 13 (2025) 18912–18936, https://doi.org/10.1109/ACCESS.2025.3532853.

[5] T. Masterman, S. Besen, M. Sawtell, A. Chao, The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: A survey, arXiv preprint arXiv:2404.11584, 2024.

[6] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, et al., The rise and potential of large language model based agents: a survey, Sci. China Inf. Sci. 68 (2) (2025) 121101.

[7] J. Xie, Z. Chen, R. Zhang, X. Wan, G. Li, Large multimodal agents: A survey, arXiv preprint arXiv:2402.15116, 2024.

[8] S. Kapoor, B. Stroebl, Z.S. Siegel, N. Nadgir, A. Narayanan, AI agents that matter, 2024, arXiv preprint arXiv:2407.01502.

[9] P. Jindal, What Are AI Agents? A Comprehensive Guide in 2025, Labellerr Blog, feb 2025, https://www.labellerr.com/blog/what-are-ai-agents-a-comprehensive-guide/, (accessed: January 2026)

[10] A. Gutowska, What are AI agents? 2025. [Online]. Available: https://www.ibm.com/think/topics/ai-agents.

[11] S. Sung, What is agentic AI? 2025. [Online]. Available: https://www.salesforce.com/agentforce/what-is-agentic-ai/.

[12] G. Marcus, The next decade in AI: four steps towards robust artificial intelligence, arXiv preprint arXiv:2002.06177, 2020.

[13] E. Sansone, R. Manhaeve, Learning symbolic representations through joint generative and discriminative training, in: NeSy-GeMs Workshop at ICLR 2023, Kigali, Rwanda, 2023. [Online]. Available: https://arxiv.org/abs/2304.11357.

[14] B.C. Colelough, W. Regli, Neuro-symbolic AI in 2024: a systematic review, in: 1st International Workshop on Logical Foundations of Neuro-Symbolic AI (LNSAI 2024), Jeju, South Korea, 2024. [Online]. Available: https://arxiv.org/abs/2501.05435.

[15] M.J. Page, D. Moher, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, et al., Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews, BMJ 372 (2021).

[16] P. Putta, E. Mills, N. Garg, S. Motwani, C. Finn, D. Garg, R. Rafailov, Agent q: Advanced reasoning and learning for autonomous AI agents, arXiv preprint arXiv:2408.07199, 2024, [Online]. Available: https://doi.org/10.48550/arXiv.2408.07199

[17] X. Zhu, Y. Chen, H. Tian, C. Tao, W. Su, C. Yang, G. Huang, B. Li, L. Lu, X. Wang, Y. Qiao, Z. Zhang, Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory, arXiv preprint arXiv:2305.17144, 2023, [Online]. Available: https://arxiv.org/pdf/2305.17144.

[18] M. Mitchell, A. Ghosh, A.S. Luccioni, G. Pistilli, Fully autonomous ai agents should not be developed, arXiv preprint arXiv:2502.02649, 2025.

[19] S.K.A. Fahad, D.W. Zhengkui, N.P. Chet, N. Wong, A.B. Ng, S. See, Advancements and applications of multimodal large language models: integration, challenges, and future directions, in: AI-Driven: Social Media Analytics and Cybersecurity, Springer, 2025, pp. 309–336.

[20] Z. Durante, Q. Huang, N. Wake, R. Gong, J.S. Park, B. Sarkar, R. Taori, Y. Noda, D. Terzopoulos, Y. Choi, et al., Agent AI: Surveying the horizons of multimodal interaction, arXiv preprint arXiv:2401.03568, 2024.

[21] A. Dingli, D. Farrugia, Neuro-Symbolic AI: Design Transparent and Trustworthy Systems that Understand the World as You Do, Packt Publishing Ltd, 2023.

[22] A. Newell, H.A. SIMON, 1 symbols and physical-symbol systems, Philos. Psychol. Contemp. Read. (2007) 407.

[23] J. Munro, A review of:rule-based expert systems: the mycin experiments of the Stanford heuristic programming project eds bruce b. Buchanan and edward h. Shortliffe addison-wesley, 1984, 748 PP, £37.00, Civ. Eng. Syst. 1 (6) (1984) 342–343.

[24] F. Rossi, P. Van Beek, T. Walsh, Handbook of Constraint Programming, Elsevier, 2006.

[25] D.B. Lenat, R.V. Guha, K. Pittman, D. Pratt, M. Shepherd, Cyc: toward programs with common sense, Commun. ACM 33 (8) (1990) 30–49.

[26] J.J. Sviokla, An examination of the impact of expert systems on the firm: the case of xcon, MIS Q. (1990) 127–140.

[27] R. Riegel, A. Gray, F. Luus, N. Khan, N. Makondo, I.Y. Akhalwaya, H. Qian, R. Fagin, F. Barahona, U. Sharma, et al., Logical neural networks, arXiv preprint arXiv:2006.13155, 2020.

[28] M. Sponner, B. Waschneck, A. Kumar, Adapting neural networks at runtime: current trends in at-runtime optimizations for deep learning, ACM Comput. Surv. 56 (10) (2024) 1–40.

[29] T.R. Besold, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K.-U. Kühnberger, L.C. Lamb, P.M.V. Lima, L. de Penning, G. Pinkas et al., Neural-symbolic learning and reasoning: a survey and interpretation 1, in: Neuro-Symbolic Artificial Intelligence: the State of the Art, IOS Press, 2021, pp. 1–51.

[30] P.C. Wason, J.S.B.T. Evans, Dual processes in reasoning? Cognition 3 (2) (1974) 141–154.

[31] J.S.B.T. Evans, K.E. Stanovich, Dual-process theories of higher cognition: advancing the debate, Perspect. Psychol. Sci. 8 (3) (2013) 223–241.

[32] D. Kahneman, Thinking, Fast and Slow, Farrar, Straus and Giroux, New York, 2011.

[33] G. Gronchi, G. Gavazzi, M.P. Viggiano, F. Giovannelli, Dual-process theory of thought and inhibitory control: an ale meta-analysis, Brain Sci. 14 (1) (2024) 101.

[34] T. Moskovitz, K.J. Miller, M. Sahani, M.M. Botvinick, Understanding dual process cognition via the minimum description length principle, PLOS Comput. Biol. 20 (10) (2024) e1012383.

[35] M. van Bekkum, M. de Boer, F. van Harmelen, et al., Modular design patterns for hybrid learning and reasoning systems, Applied Intell. 51 (9) (2021) 6528–6546.

[36] J.E. Laird, The Soar Cognitive Architecture, The MIT Press, 2012.

[37] J.E. Laird, The Soar Cognitive Architecture, MIT Press, 2019.

[38] J.R. Anderson, C.J. Lebiere, The Atomic Components of Thought, Psychology Press, New York, 1998.

[39] O. Bougzime, S. Jabbar, C. Cruz, F. Demoly, Unlocking the potential of generative ai through neuro-symbolic architectures: Benefits and limitations, arXiv preprint arXiv:2502.11269, 2025.

[40] D. Jayalath, J.B. Wendt, N. Monath, S. Tata, B. Gunel, Prism: efficient long-range reasoning with short-context llms, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025), 2025, pp. 10208–10230.

[41] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: synergizing reasoning and acting in language models, in: International Conference on Learning Representations (ICLR), 2023.

[42] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, S. Yao, Reflexion: language agents with verbal reinforcement learning, in: Advances in Neural Information Processing Systems, vol. 36, 2023.

[43] C. Michel-Delétie, M.K. Sarker, Neuro-symbolic methods for trustworthy AI: a systematic review, Neurosymbolic Artif. Intell. (2024).

[44] S. Badreddine, A.D. Garcez, L. Serafini, M. Spranger, Logic tensor networks, Artif. Intell. 303 (2022) 103649.

[45] A. Demidovskij, Automatic construction of tensor product variable binding neural networks for neural-symbolic intelligent systems, in: 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), IEEE, 2020, pp. 1–5.

[46] T. Rocktäschel, S. Riedel, Learning knowledge base inference with neural theorem provers, in: Proceedings of the 5th Workshop on Automated Knowledge Base Construction, 2016, pp. 45–50.

[47] T. Wu, M. Tjandrasuwita, Z. Wu, X. Yang, K. Liu, R. Sosic, J. Leskovec, Zeroc: a neuro-symbolic model for zero-shot concept recognition and acquisition at inference time, Adv. Neural Inf. Process. Syst. 35 (2022) 9828–9840.

[48] J. Chen, H. Li, J. Yang, Y. Liu, Q. Ai, Enhancing LLM-based agents via global planning and hierarchical execution, arXiv preprint arXiv:2504.16563, 2025, [Online]. Available: https://doi.org/10.48550/arXiv.2504.16563

[49] T.H. Trinh, Y. Wu, Q.V. Le, H. He, T. Luong, Solving olympiad geometry without human demonstrations, Nature 625 (7995) (2024) 476–482.

[50] S. Hong, M. Zhuge, J. Chen, X. Zheng, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S.K.S. Yau, Z. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, J. Schmidhuber, MetaGPT: meta programming for a multi-agent collaborative framework, in: International Conference on Learning Representations, 2024, iCLR 2024 Oral. [Online]. Available: https://arxiv.org/abs/2308.00352.

[51] H. Wei, P. Shakarian, C. Lebiere, B. Draper, N. Krishnaswamy, S. Nirenburg, Metacognitive AI: framework and the case for a neurosymbolic approach, in: International Conference on Neural-Symbolic Learning and Reasoning, Springer, 2024, pp. 60–67.

[52] P. Kouvaros, Towards formal verification of neuro-symbolic multi-agent systems, in: IJCAI, 2023, pp. 7014–7019.

[53] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N.V. Chawla, O. Wiest, X. Zhang, Large language model based multi-agents: a survey of progress and challenges, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI 2024), 2024.

[54] V. Dibia, Multi-agent LLM applications: a review of current research, tools, and challenges, Des. with AI Newsl. (Dec 2023), (Accessed 1 May 2025). [Online]. Available: https://newsletter.victordibia.com/p/multi-agent-llm-applications-a-review.

[55] J. Allen, L. Galescu, C.M. Teng, I. Perera, Conversational agents for complex collaborative tasks, AI Mag. 41 (4) (2020) 54–78.

[56] W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C.-M. Chan, H. Yu, Y. Lu, Y.-H. Hung, C. Qian, Y. Qin, X. Cong, R. Xie, Z. Liu, M. Sun, J. Zhou, Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors, in International Conference on Learning Representations, 2024, iCLR 2024 Poster. [Online]. Available: https://arxiv.org/abs/2308.10848.

[57] A.M. Sami, Z. Rasheed, K.-K. Kemell, M. Waseem, T. Kilamo, M. Saari, A.N. Duc, K. Systä, P. Abrahamsson, System for systematic literature review using multiple AI agents: Concept and an empirical evaluation, arXiv preprint arXiv:2403.08399, 2024.

[58] Z. Liu, Y. Zhang, P. Li, Y. Liu, D. Yang, Dynamic LLM-agent network: An LLM-agent collaboration framework with agent team optimization, arXiv preprint arXiv:2310.02170, 2023.

[59] B. Liu, J. Zhang, F. Lin, C. Yang, M. Peng, W. Yin, Symagent: a neural-symbolic self-learning agent framework for complex reasoning over knowledge graphs, in: Proceedings of the ACM Web Conference, 2025.

[60] Z. Xue, Z. Zhang, H. Liu, S. Yang, S. Han, Learning knowledge graph embedding with multi-granularity relational augmentation network, Expert Syst. Appl. 233 (2023) 120953.

[61] J. Mao, C. Gan, P. Kohli, J.B. Tenenbaum, J. Wu, The neuro-symbolic concept learner: interpreting scenes, words, and sentences from natural supervision, in: International Conference on Learning Representations, 2019. [Online]. Available: https://openreview.net/forum?id=H1fFfC4tPr.

[62] S. Amizadeh, H. Palangi, A. Polozov, Y. Huang, K. Koishida, Neuro-symbolic visual reasoning: disentangling, in: International Conference on Machine Learning, Pmlr, 2020, pp. 279–290.

[63] J. Xu, H. Fei, L. Pan, Q. Liu, M. Lee, W. Hsu, Faithful logical reasoning via symbolic chain-of-thought, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 13326–13365. [Online]. Available: https://aclanthology.org/2024.acl-long.720/.

[64] R.I.M. Vsevolodovna, M. Monti, Enhancing large language models through neuro-symbolic integration and ontological reasoning, arXiv preprint arXiv:2504.07640, 2025.

[65] K. Acharya, A. Velasquez, H.H. Song, A survey on symbolic knowledge distillation of large language models, IEEE Trans. Artif. Intell. 5 (12) (2024) 5928–5948, https://doi.org/10.1109/TAI.2024.3428519.

[66] H. Dong, J. Mao, T. Lin, C. Wang, L. Li, D. Zhou, Neural logic machines, in: International Conference on Learning Representations, 2019. [Online]. Available: https://openreview.net/forum?id=B1xY-hRctX.

[67] H. Shindo, M. Nishino, A. Yamamoto, Differentiable inductive logic programming for structured examples, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 5034–5041, no. 6).

[68] P. Sen, B.W.S.R. de Carvalho, R. Riegel, A. Gray, Neuro-symbolic inductive logic programming with logical neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 8212–8219, no. 8).

[69] M. Fang, S. Deng, Y. Zhang, Z. Shi, L. Chen, M. Pechenizkiy, J. Wang, Large language models are neurosymbolic reasoners, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 17985–17993, no. 16).

[70] J.J. Sun, M. Tjandrasuwita, A. Sehgal, A. Solar-Lezama, S. Chaudhuri, Y. Yue, O. Costilla Reyes, Neurosymbolic programming for science, in: NeurIPS 2022 Workshop on AI for Science, 2022. [Online]. Available: https://openreview.net/forum?id=MJtzhkSRQr.

[71] E. Marconato, G. Bontempo, E. Ficarra, S. Calderara, A. Passerini, S. Teso, Neuro-symbolic continual learning: knowledge, reasoning shortcuts and concept re-hearsal, in: Proceedings of the 40th International Conference on Machine Learning, 2023.

[72] L. Keller, D. Tanneberg, J. Peters, Neuro-symbolic imitation learning: Discovering symbolic abstractions for skill learning, arXiv preprint arXiv:2503.21406, 2025.

[73] T.X. Olausson, A. Gu, B. Lipkin, C.E. Zhang, A. Solar-Lezama, J.B. Tenenbaum, R.P. Levy, Linc: a neurosymbolic approach for logical reasoning by combining language models with first-order logic provers, in The 2023 Conference on Empirical Methods in Natural Language Processing.

[74] J. Rabold, A neural-symbolic approach for explanation generation based on sub-concept detection: an application of metric learning for low-time-budget labeling, KI-Künstl. Intell. 36 (3) (2022) 225–235.

[75] K. Sanders, N. Weir, B. Van Durme, Tv-trees: multimodal entailment trees for neuro-symbolic video reasoning, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), 2024.

[76] Z. Li, Y. Huang, Z. Li, Y. Yao, J. Xu, T. Chen, X. Ma, J. Lu, Neuro-symbolic learning yielding logical constraints, Adv. Neural Inf. Process. Syst. 36 (2023) 21635–21657.

[77] S.H. Bach, M. Broecheler, B. Huang, L. Getoor, Hinge-loss markov random fields and probabilistic soft logic, J. Mach. Learn. Res. 18 (109) (2017) 1–67, [Online]. Available: http://jmlr.org/papers/v18/15-631.html.

[78] M. Gaur, A. Sheth, Building trustworthy neurosymbolic AI systems: consistency, reliability, explainability, and safety, AI Mag. 45 (1) (2024) 139–155.

[79] S. Mehrotra, C.C. Jorge, C.M. Jonker, M.L. Tielman, Integrity-based explanations for fostering appropriate trust in AI agents, ACM Trans. Interact. Intell. Syst. 14 (1) (2024) 1–36.

[80] L. DeLong, Y. Gadiya, J.D. Fleuriot, D. Domingo-Fernández, Neurosymbolic AI reveals biases and limitations in ML-driven drug discovery, in: AI4D3 Workshop at NeurIPS, 2023, poster.

[81] A. Gomaa, M. Feld, Towards adaptive user-centered neuro-symbolic learning for multimodal interaction with autonomous systems, in: Proceedings of the 25th International Conference on Multimodal Interaction, 2023, pp. 689–694.

[82] Q. Lu, R. Li, E. Sagheb, A. Wen, J. Wang, L. Wang, J.W. Fan, H. Liu, Explainable diagnosis prediction through neuro-symbolic integration, in: AMIA Joint Summits on Translational Science Proceedings, 2025, pp. 332–341.

[83] R.J. Tong, X. Hu, Future of education with neuro-symbolic AI agents in self-improving adaptive instructional systems, Front. Digit. Educ. 1 (2) (2024) 198–212.

[84] A. Capitanelli, F. Mastrogiovanni, A framework for neurosymbolic robot action planning using large language models, Front. Neurorobot. 18 (2024) 1342786.

[85] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al., Do As I Can, Not As I Say: Grounding language in robotic affordances, arXiv preprint arXiv:2204.01691, 2022.

[86] Z. Li, Y. Yao, T. Chen, J. Xu, C. Cao, X. Ma, J. Lü, Softened symbol grounding for neuro-symbolic systems, in: International Conference on Learning Representations, 2023. [Online]. Available: https://openreview.net/forum?id=X0aVQw1QpG.

[87] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q.V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Adv. Neural Inf. Process. Syst. 35 (2022) 24824–24837.

[88] J. Li, G. Li, Y. Li, Z. Jin, Structured chain-of-thought prompting for code generation, ACM Trans. Softw. Eng. Methodol. 34 (2) (2025) 1–23.

[89] S. Yao, D. Yu, J. Zhao, I. Shafran, T.L. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: deliberate problem solving with large language models, in: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023), 2023.

[90] X. Wang, J. Wei, D. Schuurmans, Q.V. Le, E.H. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, in: International Conference on Learning Representations, 2023, iCLR 2023.

[91] F. Rezazadeh, H. Chergui, M. Debbah, H. Song, D. Niyato, L. Liu, Agentic world modeling for 6G: Near-real-time generative state-space reasoning, 2025, [Online]. Available: https://arxiv.org/abs/2511.02748.

[92] Q. Pan, W. Ji, Y. Ding, J. Li, S. Chen, J. Wang, J. Zhou, Q. Chen, M. Zhang, Y. Wu, L. He, A survey of slow thinking-based reasoning llms using reinforcement learning and test-time scaling law, Inf. Process. Manag. 63 (2) (2026) 104394.

[93] C. Snell, J. Lee, K. Xu, A. Kumar, Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning, in: Proceedings of the 2025 International Conference on Learning Representations (ICLR 2025), 2025.

[94] S. Schmidgall, R. Ziaei, C. Harris, E. Reis, J. Jopling, M. Moor, Agentclinic: a multimodal agent benchmark to evaluate AI in simulated clinical environments, arXiv preprint arXiv:2405.07960, 2024.

[95] S. Järvelä, A. Nguyen, A. Hadwin, Human and artificial intelligence collaboration for socially shared regulation in learning, Br. J. Educ. Technol. 54 (5) (2023) 1057–1076.

[96] C.D. Jaldi, E. Ilkou, N. Schroeder, C. Shimizu, Education in the era of neurosymbolic AI, J. Web Semant. 85 (2025) 100857.

[97] K. Vyas, D. Graux, S. Montella, P. Vougiouklis, R. Lai, K. Li, Y. Ren, J.Z. Pan, An extensive evaluation of PDDL capabilities in off-the-shelf LLMs, CoRR, arXiv:2502.20175, 2025, [Online]. Available: https://doi.org/10.48550/arXiv.2502.20175

[98] M. Fox, D. Long, PDDL2. 1: an extension to PDDL for expressing temporal planning domains, J. Artif. Intell. Res. 20 (2003) 61–124.

[99] K. Erol, J.A. Hendler, D.S. Nau, UMCP: a sound and complete procedure for hierarchical task-network planning, in: Aips, vol. 94, 1994, pp. 249–254.

[100] T. Birr, C. Pohl, A. Younes, T. Asfour, AutoGPT+P: affordance-based task planning with large language models, in: Robotics: Science and Systems, 2024.

[101] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: language models can teach themselves to use tools, Adv. Neural Inf. Process. Syst. 36 (2023) 68539–68551.

[102] N. Dziri, X. Lu, M. Sclar, X.L. Li, L. Jiang, B.Y. Lin, S. Welleck, P. West, C. Bhagavatula, R. Le Bras, et al., Faith and fate: limits of transformers on compositionality, Adv. Neural Inf. Process. Syst. 36 (2023) 70293–70332.

[103] J. Liu, C. Yu, J. Gao, Y. Xie, Q. Liao, Y. Wu, Y. Wang, LLM-powered hierarchical language agent for real-time human-AI coordination, in: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS '24), Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2024, pp. 1219–1228.

[104] B. Xu, Z. Peng, B. Lei, S. Mukherjee, Y. Liu, D. Xu, Rewoo: Decoupling reasoning from observations for efficient augmented language models, arXiv preprint arXiv:2305.18323, 2023.

[105] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, arXiv preprint arXiv:2501.12948, 2025, [Online]. Available: https://arxiv.org/abs/2501.12948.

[106] M. Hu, P. Zhao, C. Xu, Q. Sun, J.-G. Lou, Q. Lin, P. Luo, S. Rajmohan, Agentgen: enhancing planning abilities for large language model based agents via environment and task generation, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2025), 2025.

[107] Z. Wang, K. Wang, Q. Wang, P. Zhang, L. Li, Z. Yang, K. Yu, M.N. Nguyen, L. Liu, E. Gottlieb, M. Lam, Y. Lu, K. Cho, J. Wu, L. Fei-Fei, L. Wang, Y. Choi, M. Li, Ragen: understanding self-evolution in LLM agents via multi-turn reinforcement learning, 2025. [Online]. Available: https://arxiv.org/abs/2504.20073.

[108] A. Novikov, N. Vũ, M. Eisenberger, E. Dupont, P.-S. Huang, A.Z. Wagner, S. Shirobokov, B. Kozlovskii, F.J.R. Ruiz, A. Mehrabian, et al., AlphaEvolve: A coding agent for scientific and algorithmic discovery, arXiv preprint arXiv:2506.13131, 2025.

[109] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, Adv. Neural Inf. Process. Syst. 33 (2020) 9459–9474.

[110] R. Friel, M. Belyi, A. Sanyal, Ragbench: Explainable benchmark for retrieval-augmented generation systems, arXiv preprint arXiv:2407.11005, 2024.
[111] A. Huet, Z. Ben Houidi, D. Rossi, Episodic memories generation and evaluation benchmark for large language models, in: International Conference on Learning Representations, 2025, iCLR 2025 Poster. [Online]. Available: https://arxiv.org/abs/2501.13121.
[112] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, A. Awadallah, R.W. White, D. Burger, C. Wang, Autogen: enabling next-gen LLM applications via multi-agent conversation, in: Conference on Language Modeling (COLM), 2024.
[113] G. Li, H. Hammoud, H. Itani, D. Khizbullin, B. Ghanem, "Camel: communicative agents for "mind" exploration of large language model society,", Adv. Neural Inf. Process. Syst. 36 (2023) 51991–52008.
[114] C. Li, D. Teney, L. Yang, Q. Wen, X. Xie, J. Wang, Culturepark: boosting cross-cultural understanding in large language models, Adv. Neural Inf. Process. Syst. 37 (2024) 65183–65216.
[115] X. Zhou, H. Zhu, L. Mathur, R. Zhang, H. Yu, Z. Qi, L.-P. Morency, Y. Bisk, D. Fried, G. Neubig, M. Sap, Sotopia: interactive evaluation for social intelligence in language agents, in: International Conference on Learning Representations, 2024, iCLR 2024 Spotlight. [Online]. Available: https://arxiv.org/abs/2310.11667.
[116] N. Papoulias, Neuroql: A neuro-symbolic language and dataset for inter-subjective reasoning, arXiv preprint arXiv:2303.07146, 2023.
[117] J.P. Inala, Y. Yang, J. Paulos, Y. Pu, O. Bastani, V. Kumar, M. Rinard, A. Solar-Lezama, Neurosymbolic transformers for multi-agent communication, Adv. Neural Inf. Process. Syst. 33 (2020) 13597–13608.
[118] X. Yi, Z. Zhou, C. Cao, Q. Niu, B. Han, Towards efficient and scalable multi-agent reasoning via Bayesian nash equilibrium, 2025. [Online]. Available: https://openreview.net/forum?id=MWSoYGPexK.
[119] C. Amato, An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning, arXiv preprint arXiv:2409.03052, 2024.
[120] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C.D. Manning, Hotpotqa: a dataset for diverse, explainable multi-hop question answering, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2369–2380.
[121] Q.-H. Vuong, M.-T. Ho, The disruptive alphageometry: is it the beginning of the end of Mathematics education? AI & SOCIETY (2024) 1–3.
[122] F. Fabiano, V. Pallagani, M.B. Ganapini, L. Horesh, A. Loreggia, K. Murugesan, F. Rossi, B. Srivastava, Plan-SOFAI: a neuro-symbolic planning architecture, in: Neuro-Symbolic Learning and Reasoning in the Era of Large Language Models, 2023.
[123] A. Didolkar, A. Goyal, N.R. Ke, S. Guo, M. Valko, T. Lillicrap, D. Jimenez Rezende, Y. Bengio, M.C. Mozer, S. Arora, Metacognitive capabilities of LLMs: an exploration in mathematical problem solving, Adv. Neural Inf. Process. Syst. 37 (2024) 19783–19812.
[124] P. Shakarian, H. Wei (Eds.), Metacognitive Artificial Intelligence, Cambridge University Press, 2025.
[125] X. Zhang, V.S. Sheng, Neuro-symbolic AI: Explainability, challenges, and future trends, arXiv preprint arXiv:2411.04383, 2024.
[126] J. Ott, A. Ledaguenel, C. Hudelot, M. Hartwig, How to think about benchmarking neurosymbolic AI? in: 17th International Workshop on Neural-Symbolic Learning and Reasoning-NESY 2023, 2023.
[127] C. Shengyuan, Y. Cai, H. Fang, X. Huang, M. Sun, Differentiable neuro-symbolic reasoning on large-scale knowledge graphs, Adv. Neural Inf. Process. Syst. 36 (2023) 28139–28154.
[128] S. I. for Human-Centered Artificial Intelligence, The 2025 AI index report, 2025.
[129] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489.
[130] S. I. for Human-Centered Artificial Intelligence, The 2024 AI index report, 2024.
[131] L.N. DeLong, R. Fernández Mir, J.D. Fleuriot, Neurosymbolic AI for reasoning over knowledge graphs: a survey, IEEE Trans. Neural Netw. Learn. Syst. 36 (5) (2025) 7822–7842.
[132] Z. Lu, I. Afridi, H.J. Kang, I. Ruchkin, X. Zheng, Surveying neuro-symbolic approaches for reliable artificial intelligence of things, J. Rel. Intell. Environ. 10 (3) (2024) 257–279.
[133] P. Barbiero, F. Giannini, G. Ciravegna, M. Diligenti, G. Marra, Relational concept bottleneck models, Adv. Neural Inf. Process. Syst. 37 (2024) 77663–77685.
[134] B. Rao, W. Eiers, C. Lipizzi, Neural theorem proving: Generating and structuring proofs for formal verification, arXiv preprint arXiv:2504.17017, 2025.
[135] H.U. Sheikh, S. Khadka, S. Miret, S. Majumdar, M. Phielipp, Learning intrinsic symbolic rewards in reinforcement learning, in: 2022 International Joint Conference on Neural Networks (IJCNN), IEEE, 2022, pp. 1–8.
[136] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, S. Zhang, X. Deng, A. Zeng, Z. Du, C. Zhang, S. Shen, T. Zhang, Y. Su, H. Sun, M. Huang, et al., Agentbench: evaluating LLMs as agents, in: International Conference on Learning Representations, 2024, iCLR 2024 Poster. [Online]. Available: https://arxiv.org/abs/2308.03688.
[137] B.Y. Lin, Y. Deng, K. Chandu, A. Ravichander, V. Pyatkin, N. Dziri, R. Le Bras, Y. Choi, Wildbench: benchmarking LLMs with challenging tasks from real users in the wild, in: International Conference on Learning Representations, 2025, iCLR 2025 Spotlight.
[138] M. Geva, D. Khashabi, E. Segal, T. Khot, D. Roth, J. Berant, Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies, 2021, https://huggingface.co/datasets/voidful/StrategyQA (accessed: 30 April 2025).
[139] S. Harnad, The symbol grounding problem, Phys. D: Nonlinear Phenom. 42 (1–3) (1990) 335–346.
[140] M. Taddeo, L. Floridi, Solving the symbol grounding problem: a critical review of fifteen years of research, J. Exp. Theor. Artif. Intell. 17 (4) (2005) 419–445.
[141] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, A. Anandkumar, Voyager: an open-ended embodied agent with large language models, Trans. Mach. Learn. Res. (2024), accepted by TMLR. [Online]. Available: https://voyager.minedojo.org.
[142] A. Sheth, V. Pallagani, K. Roy, Neurosymbolic AI for enhancing instructability in generative AI, IEEE Intell. Syst. 39 (5) (2024) 5–11.
[143] S.G.B. Johnson, A.-H. Karimi, Y. Bengio, N. Chater, T. Gerstenberg, K. Larson, S. Levine, M. Mitchell, I. Rahwan, B. Schölkopf, I. Grossmann, Imagining and building wise machines: The centrality of AI metacognition, CoRR, arXiv:2411.02478, 2024, revised 2025. [Online]. Available: https://arxiv.org/abs/2411.02478.
[144] L. Narens, A. Graf, T.O. Nelson, Metacognitive aspects of implicit/explicit memory, in: Implicit Memory and Metacognition, Psychology Press, 2014, pp. 137–170.
[145] R. Ackerman, V.A. Thompson, Meta-reasoning: monitoring and control of thinking and reasoning, Trends Cogn. Sci. 21 (8) (2017) 607–617.
[146] P. Lindes, The common model of cognition and humanlike language comprehension, Procedia Comput. Sci. 145 (2018) 765–772.
[147] Tech Dose, Training AI agents that don't fall apart: how ragen and starpo are solving the 'echo trap', Apr, 2025, https://medium.com/@tech.dose/training-ai-agents-that-dont-fall-apart-how-ragen-and-starpo-are-solving-the-echo-trap-c149b23248fa.
[148] G. Wang, W. Wu, G. Ye, Z. Cheng, X. Chen, H. Zheng, Decoupling metacognition from cognition: a framework for quantifying metacognitive ability in LLMs, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, 2025, pp. 25353–25361, no. 24).
[149] S. Bortolotti, E. Marconato, T. Carraro, P. Morettin, E. van Krieken, A. Vergari, S. Teso, A. Passerini, A neuro-symbolic benchmark suite for concept quality and reasoning shortcuts, in: NeurIPS 2024, 2024.
[150] Arize AI Team, Libreeval: Phoenix open source hallucination evaluation model & dataset (version 1.1), 2025, https://arize.com/llm-hallucination-dataset/ (accessed: 23 November 2025).
[151] N. Ngu, A. Taparia, G.I. Simari, M.A. Leiva, R. Senanayake, P. Shakarian, N.D. Bastian, J. Corcoran, Multiple distribution shift – aerial (MDS-A): a dataset for test-time error detection and model adaptation, in: Proceedings of the AAAI Spring Symposium, 2025, pp. 379–383.
[152] J. Renkhoff, K. Feng, M. Meier-Doernberg, A. Velasquez, H.H. Song, A survey on verification and validation, testing and evaluations of neurosymbolic artificial intelligence, IEEE Trans. Artif. Intell. 5 (8) (2024) 3765–3779.
[153] W. Choi, J. Park, S. Ahn, D. Lee, H. Woo, Nesyc: a neuro-symbolic continual learner for complex embodied tasks in open domains, in: International Conference on Learning Representations, 2025. [Online]. Available: https://openreview.net/forum?id=VoayJihXra.
[154] Y. Chervonyi, T.H. Trinh, M. Olšák, X. Yang, H.H. Nguyen, M. Menegali, J. Jung, J. Kim, V. Verma, Q.V. Le, et al., Gold-medalist performance in solving olympiad geometry with alphageometry2, J. Mach. Learn. Res. 26 (241) (2025) 1–39.
[155] B. Liu, X. Li, J. Zhang, J. Wang, T. He, S. Hong, H. Liu, S. Zhang, K. Song, K. Zhu, et al., Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems, arXiv preprint arXiv:2504.01990, 2025.
[156] B.P. Bhuyan, A. Ramdane-Cherif, T.P. Singh, R. Tomar, Common sense reasoning for neuro-symbolic AI, in: Neuro-Symbolic Artificial Intelligence: Bridging Logic and Learning, Springer, 2024, pp. 271–290.
[157] J. Huang, Z. Li, B. Chen, K. Samel, M. Naik, L. Song, X. Si, Scallop: from probabilistic deductive databases to scalable differentiable reasoning, Adv. Neural Inf. Process. Syst. 34 (2021) 25134–25145.
[158] L. Hammond, A. Chan, J. Clifton, J. Hoelscher-Obermaier, A. Khan, E. McLean, C. Smith, W. Barfuss, J.N. Foerster, T. Gavenciak, T.A. Han, E. Hughes, V. Kovarík, J. Kulveit, J.Z. Leibo, C. Oesterheld, C. Schröder de Witt, N. Shah, M.P. Wellman, P. Bova, et al., Multi-agent risks from advanced AI, CoRR, arXiv:2502.14143, 2025, revised 2025. [Online]. Available: https://arxiv.org/abs/2502.14143.