Governed Hybrid Agency: A Critical Evidence Map of Cognitive Architectures for Foundation-Model Agents

DARIUS SABALIAUSKAS*, Cybersocial Systems Engineering Group, Institute of Data Science and Digital Technologies, Faculty of Mathematics and Informatics, Vilnius University, Lithuania
AISTE DIRZYTE, Faculty of Creative Industries, Vilnius Gediminas Technical University, Lithuania and Institute of Psychology, Mykolas Romeris University, Lithuania

Background. Cognitive architectures specify the persistent memories, representations, learning processes, and control loops that allow an artificial agent to remain coherent across tasks and time. Foundation models provide broad generative competence, but do not by themselves ensure grounded action, reliable long-horizon memory, calibrated uncertainty, execution monitoring, or accountable learning.
Objectives. This critical integrative review examines how cognitive-architecture principles are being used to govern foundation-model agents. Taking the last comprehensive pre-foundation-model survey of cognitive architectures (Kotseruba & Tsotsos, first published online in 2018) as its historical baseline, it maps developments across twelve ability channels, identifies recurring integration patterns, and derives testable design and evaluation requirements.
Methods. We combined a historical-theoretical synthesis with a focused, PRISMA-informed evidence map of English-language records indexed in Scopus and the Web of Science Core Collection, searched in July 2026 over the 2018–2026 window. Queries required terms for cognitive architectures or architecture-level cognitive systems, foundation or generative models, and governance-relevant mechanisms. The searches identified 1,173 records; after deduplication, rule-based filtering, and LLM-assisted title-and-abstract assessment, 119 works were retained and examined in full text. Because the 660 candidate records were not all independently human-rescreened, the corpus is interpreted as a reproducible focused evidence map rather than an exhaustive systematic review.
Results. All included works were published or available online in 2023–2026. Five recurring integration patterns were identified: foundation models used for knowledge or generative memory, language or multimodal interfaces, heuristic proposal or planning, rule or model bootstrapping, and bidirectional neuro-symbolic coupling. Across these patterns, a recurring design principle is to place foundation models inside persistent control loops that provide explicit state, provenance-aware memory, verification, metacognitive arbitration, execution monitoring, and recovery. Evaluation remains uneven, particularly for calibration, retention, transfer, long-horizon robustness, disturbance recovery, and reproducibility.
Conclusions. We propose governed hybrid agency, an Architectural Governance Stack, typed interface contracts, and a risk-based evaluation scorecard. These are integrative, falsifiable comparison targets rather than a validated universal architecture, and their value should be tested through matched ablations, intervention-sensitive evaluation, and cross-domain replication. A worked example instantiates the framework as a modular requirements-engineering reasoning architecture.

JAIR Track: Surveys

JAIR Associate Editor: Roberta Calegari

*Corresponding author.

Authors' Contact Information: Darius Sabaliauskas, ORCID: 0009-0007-0371-3435, darius.sabaliauskas@mif.vu.lt, Cybersocial Systems Engineering Group, Institute of Data Science and Digital Technologies, Faculty of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania; Aiste Dirzyte, ORCID: 0000-0003-2057-3108, aiste.dirzyte@vilniustech.lt, Faculty of Creative Industries, Vilnius Gediminas Technical University, Vilnius, Lithuania and Institute of Psychology, Mykolas Romeris University, Vilnius, Lithuania.



This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2026 Copyright held by the owner/author(s).
DOI: 10.1613/jair.1.23727

JAIR Reference Format:
Darius Sabaliauskas and Aiste Dirzyte. 2026. Governed Hybrid Agency: A Critical Evidence Map of Cognitive Architectures for Foundation-Model Agents. *Journal of Artificial Intelligence Research* 87, Article 22 (November 2026), 47 pages. DOI: 10.1613/jair.1.23727

1 Introduction

This section motivates the renewed role of cognitive architectures in the foundation-model era, positions the review against prior surveys, and states the review questions and contributions.

1.1 Why Cognitive Architectures Matter after Foundation Models

Cognitive architectures (CAs) are computational commitments about the relatively stable structures that make cognition possible: the memories an agent maintains, the representations it manipulates, the processes that select and sequence actions, the learning mechanisms that alter future behavior, and the control loops that coordinate these functions. They are therefore more than collections of algorithms. A CA specifies how components communicate over time and how an agent remains coherent across changing tasks and environments (Kotseruba and Tsotsos 2020; Lieto et al. 2018). The Common Model of Cognition (CMC), and established systems such as Soar, ACT-R, CLARION, ICARUS, LIDA, and Sigma, exemplify the field's long-standing effort to connect cognitive theory, artificial intelligence, and practical system design (Laird 2012; Lebiere, Cranford, et al. 2022).

The foundation-model era changes the engineering context but not the architectural problem. Large language models (LLMs), vision-language models, and related learned policies provide remarkable language, perception, and generative capabilities. Yet a model that can propose a plausible next utterance or action does not by itself supply persistent episodic memory, explicit task state, calibrated uncertainty, verifiable planning, recovery from execution failure, or principled control over when to use slow rather than fast cognition. Language-agent frameworks such as CoALA already rediscover modular memory, structured action spaces, and decision cycles familiar from cognitive architectures (Sumers et al. 2024). Empirical integrations similarly show that LLMs can improve interaction, knowledge acquisition, and proposal generation, while classical planners or cognitive control mechanisms often remain superior in reliability and execution quality (González-Santamarta, Rodríguez-Lera, Fernández-Llamas, et al. 2024; Zhu and Simmons 2024).

The central issue is thus no longer whether cognitive architectures will be displaced by foundation models. It is how architectures should govern, constrain, remember, verify, and learn from such models. Recent work increasingly treats the LLM as one component - a noisy knowledge source, generative memory, language interface, heuristic proposer, or rule generator - within a larger system (Dancy and Workman 2024; Joshi and Ustun 2024; Kirk et al. 2024; O. J. Romero et al. 2023; Wray et al. 2025). This shift makes the architecture-model interface a first-class scientific and practical object.

1.2 Why an Updated Integrative Review Is Needed

Kotseruba and Tsotsos (2020; first published online in 2018, the temporal baseline adopted here) surveyed 84 cognitive architectures and more than 900 practical projects, establishing the most influential map of the field. Their review showed broad architectural diversity but limited integration across cognitive abilities, and it noted that deep learning had not yet been systematically incorporated into unified CA frameworks. The longer historical trajectory, subsequently synthesized by Kotseruba and Tsotsos (2025), shows that architectures do not evolve through simple replacement: symbolic, emergent, hybrid, embodied, probabilistic, and generative traditions accumulate and are recombined. Since 2018, the landscape has nevertheless changed along several axes: hybrid symbolic-subsymbolic systems were the majority approach in a focused review of autonomous-robotics architectures (González-Santamarta, Rodríguez-Lera, and Matellán-Olivera 2025); ROS and ROS 2 have become common implementation substrates; LLMs have entered planning, memory, communication, and learning loops;

predictive processing and active inference have become implementable design frameworks; and metacognition has moved toward explicit architectural specification (Ciria et al. 2021; González-Santamarta, Rodríguez-Lera, and Matellán-Olivera 2025; Laird, Lebiere, Rosenbloom, and Stocco 2026; Ororbia and Kelly 2024; Ukov and Tsochev 2025).

Existing reviews illuminate parts of this transition. González-Santamarta, Rodríguez-Lera, and Matellán-Olivera (2025) focus on behavior generation in autonomous robotics. Neuro-symbolic surveys classify integration mechanisms but do not organize them by the cognitive abilities they implement (Acharya and Song 2025; Wan et al. 2025; W. Wang and Yang 2024). Agentic-AI surveys describe current orchestration frameworks but often detach them from the accumulated design knowledge of cognitive science and classical architectures (Abou Ali and Dornaika 2025; Xi et al. 2025). The 2025 monograph by Kotseruba and Tsotsos provides a broad historical and computational account, but a concise evidence map connecting post-2018 ability shifts, theoretical frameworks, LLM integration roles, implementation maturity, and practitioner choices remains useful. Broader surveys track the shift from classical multi-agent systems to foundation-model-enabled agents and the augmentation of large language models with cognitive functions (Navarro 2025; Z. Wang, Gong, et al. 2026).

1.3 Review Questions and Contributions

Building on the core abilities and adjacent capability categories in the 2018 survey, the present review uses twelve analytic channels: perception, attention, action selection, memory, learning, reasoning, planning, metacognition, communication, motivation/emotion, social cognition, and neuro-symbolic integration. Strictly, these comprise eleven agent-level abilities plus one cross-cutting integration dimension: neuro-symbolic integration is a way of combining representations rather than a stand-alone ability, and it is treated as such throughout (we retain the shorthand "twelve channels" for readability). Four questions guide the synthesis:

RQ1. Which architecture-level commitments recur when foundation models are integrated into persistent cognitive systems?

RQ2. How are these commitments distributed across the twelve ability channels — perception, attention, action selection, memory, learning, reasoning, planning, metacognition, communication, motivation/emotion, social cognition, and neuro-symbolic integration?

RQ3. Which roles are assigned to foundation models, and what benefits and failure modes are reported for each role?

RQ4. Which evidence and evaluation practices are needed to distinguish functional coverage from reliable, governable agency?

The article makes four contributions. First, it provides a focused evidence map of foundation-model-related cognitive-architecture work and separates corpus-derived findings from broader theoretical context, offering an ability-by-ability map rather than another architecture catalogue (with two synthesis tables pairing each ability channel and each theoretical framework with the systems that implement it, Tables S1 and S2). Second, it synthesizes five recurring integration patterns and maps them to twelve ability channels, implementation evidence, and governance requirements. Third, it proposes governed hybrid agency and an Architectural Governance Stack as falsifiable system-level comparison targets, including typed interfaces for state, provenance, uncertainty, authorization, and veto. Fourth, it translates the synthesis into a risk-based evaluation scorecard and a concrete experimental agenda. The intended value is both scientific — clarifying mechanisms, boundary conditions, and comparison targets — and practical — helping developers choose components, interfaces, and metrics appropriate to task risk and deployment conditions. To show how the framework's governance commitments map onto concrete architectures, the review includes an illustrative worked instantiation — a modular requirements-engineering reasoning model — together with a criteria-based, auditable rubric for selecting and revising the

Table 1. Reader's map: where each review question is answered and which artefacts summarize the answer.

Question or component	Sections	Key artefacts
RQ1: Recurring architecture-level commitments	2.11, 5.1, 5.4	Table 2; Figure 2; Table 6 (derivation of the stack)
RQ2: Distribution across the twelve ability channels	4.1, 4.2, 4.5	Table 5 (coverage matrix); Tables S1–S3
RQ3: Roles of foundation models, benefits, and failure modes	4.4, 5.4	Table 4 (integration patterns); Table 8 (decision guide)
RQ4: Evidence and evaluation practices	4.7, 6	Table 10 (scorecard); hypotheses H1–H7
Positioning against existing frameworks	5.2	Table 7
Worked example	5.5	Table 9
Screening protocol	3.3, 3.6	Figure 1; Table 3

architecture assigned to each reasoning module; it is presented as an illustration, not as evidence that the proposed stack has already been validated.

What is new relative to existing frameworks can be stated precisely. Reference architectures such as the CMC specify the functional organization of cognition but were formulated before foundation models and assign them no specific role; language-agent frameworks such as CoALA describe memory modules, action spaces, and a propose–evaluate–select decision cycle around a language model, but leave open what typing, provenance, and authorization a model output must carry before it is written to memory or executed. Governed hybrid agency makes that interface the organizing principle. The foundation model is treated as a fallible generative service, and the framework specifies what must happen before its outputs are admitted: typed messages that carry provenance, uncertainty, and permitted use; an explicit verification and veto path; and a metacognitive plane that decides when to trust, verify, abstain, or escalate. The framework also differs in how it is derived and tested: each layer of the stack is traced to recurring failure modes in the reviewed corpus (Table 6) rather than stipulated, and each commitment is paired with a discriminating test (Tables 2 and 10), so that it can be falsified component by component. Section 5.2 and Table 7 set out the comparison with the CMC, CoALA, BDI, autonomic MAPE-K loops, and SOFAI in detail.

The review addresses two constituencies within the applied AI community: researchers and practitioners. Practitioners deploying AI systems in robotics, manufacturing, healthcare, cyberdefense, personalization, and education may proceed directly to Table S4 (domain-specific value propositions), Table 8 (a risk-based decision guide for selecting an LLM-CA integration pattern), and Table 10 (a minimum evaluation scorecard). Researchers will find the discriminating tests in Table 2, the eight propositions in Section 2.11, and the falsifiable hypotheses in Section 6.2 most useful as comparison targets for new systems.

Table 1 maps each review question to the sections and artefacts that answer it.

2 Theoretical Framework: From Unified Cognition to Governed Hybrid Agency

This section develops the integrative theoretical framework: it traces the evolution of cognitive architectures, treats cognitive abilities as architectural commitments, and synthesizes the twelve ability channels into the governed-hybrid-agency account.

2.1 The Evolution of Cognitive Architectures

Cognitive architectures emerged from the ambition to formulate unified theories of cognition: computational accounts in which memory, perception, action, learning, and problem solving are not isolated models but coordinated parts of one temporally extended system (Newell 1990). Production-system architectures such as Soar and ACT-R made this commitment explicit through working-memory states, long-term declarative and procedural knowledge, and recurrent decision cycles (Anderson 2007; Laird 2012). Their scientific value lay not only in performance, but in the possibility of tracing behavior to stable mechanisms and comparing model predictions with human data.

The field subsequently diversified rather than following a single replacement path. Emergent and connectionist systems emphasized distributed representation and adaptation; hybrid architectures such as CLARION combined implicit and explicit knowledge; global-workspace and embodied systems emphasized recurrent broadcasting, perception-action coupling, and situated control; and probabilistic approaches introduced generative models, Bayesian inference, and uncertainty. The CMC consolidated a shared functional abstraction across several lineages, while the broader historical evidence shows persistent pluralism in representation and process (Kotseruba and Tsotsos 2020; Laird, Lebiere, and Rosenbloom 2017).

Foundation models create a fourth phase. Large learned models can supply language competence, multimodal pattern completion, broad prior knowledge, and heuristic action proposals at a scale unavailable to earlier architectures (Sumers et al. 2024; Xi et al. 2025). However, scale relocates rather than dissolves the architecture problem. A model call does not determine which information becomes a belief, how an episode is retained, when a slow solver should override a fast proposal (Fabiano et al. 2024), how uncertainty crosses module boundaries (Friston et al. 2017), or how the agent remains accountable across time (T. Liu and van der Schaar 2025). The defining evolution is therefore from architectures as self-contained implementations of cognition toward architectures as governance systems for heterogeneous learned and symbolic resources (Laird, Wray, et al. 2024; Wray et al. 2025). This article terms the resulting target governed hybrid agency.

2.2 Cognitive Abilities as Architectural Commitments

The twelve ability channels used in this review (Kotseruba and Tsotsos 2020) are analytical lenses, not a claim that the mind or an artificial agent contains twelve encapsulated modules. An ability may be localized, distributed, or emergent from several interacting mechanisms. Perception depends on attention and memory; planning depends on reasoning, action selection, and world models; metacognition depends on representations of lower-level processes (Nelson and Narens 1990); and social cognition recruits perception, memory, language, and inference (Z. Wang, Davis, et al. 2025). The relevant architectural question is therefore not whether an ability name appears in a diagram, but what representational and causal commitments make the ability possible.

For each ability, the framework applies four tests. First, the representation test asks what states the system explicitly maintains: objects, goals, episodes, confidence, other agents, or symbolic constraints. Second, the process test asks what operation changes those states: prediction-error minimization, production firing, retrieval, search, Bayesian inversion, or learned policy execution (Anderson 2007; Friston et al. 2017; Laird 2012). Third, the coordination test asks how the process competes or cooperates with other processes across time. Fourth, the evidence test asks what intervention, ablation, generalization, or calibration result would distinguish the proposed mechanism from a simpler alternative. These tests separate functional coverage from cognitive metaphor and provide a common language for symbolic, neural, probabilistic, and hybrid implementations.

2.3 The Common Model of Cognition as a Stable Reference Architecture

The CMC is the strongest cross-lineage answer to the question of what remains invariant across mature cognitive architectures. It abstracts a recurrent cognitive cycle linking perception and action to a capacity-limited working

memory, procedural action selection, and longer-term declarative and procedural memories (Laird, Lebiere, and Rosenbloom 2017). It is deliberately less specific than ACT-R, Soar, or Sigma: it constrains functional organization and communication while allowing different algorithms and representational substrates. This makes it a reference architecture rather than a single executable system.

Post-2018 work has developed the CMC in four directions. Neuroimaging studies have tested whether its predicted large-scale communication structure is consistent with multi-task brain data, including bottom-up Granger-causal reconstruction that suggests an expanded role for episodic memory (Hake et al. 2022). CogNGen recasts the model in a predictive-processing and hyperdimensional substrate with local Hebbian adaptation (Ororbia and Kelly 2022). West et al. (2024) place generative networks in peripheral shadow production systems whose outputs are interpreted by a central production system. Laird, Lebiere, Rosenbloom, and Stocco (2026) add metacognition by allowing working memory to represent the agent's own capabilities, processes, and outcomes, so that ordinary reasoning mechanisms can operate on self-representations.

The CMC contributes a crucial boundary condition to foundation-model integration. Replacing a language, perception, or retrieval component does not by itself replace the architecture (West et al. 2024), because the architecture resides in the persistent organization of state, memory, selection, learning, and communication (Laird, Lebiere, and Rosenbloom 2017). Conversely, conformity to the CMC does not guarantee intelligence or cognitive fidelity. Empirical claims must still identify the implemented process, its parameters, and the behavior that depends causally on it.

2.4 Predictive Processing and Active Inference as a Cross-Cutting Process Family

Predictive processing treats perception as inference under a hierarchical generative model. Higher levels predict lower-level states; mismatches produce prediction errors; and precision determines how strongly particular errors update beliefs. Active inference extends this logic to action by selecting policies that reduce expected free energy, combining pragmatic value (reaching preferred states) with epistemic value (reducing uncertainty) (Ciria et al. 2021; Friston et al. 2017). The family offers a common vocabulary for perception, attention, action, learning, and curiosity, thereby challenging strictly feed-forward or modular input-output accounts.

Architecturally, a strong predictive-processing commitment requires more than a learned predictor. The system must maintain a generative model that links hidden states, observations, and actions; represent uncertainty or precision; update beliefs from prediction errors; and expose how policies change expected observations. CogNGen is a prominent attempt to realize these commitments through neural generative coding, local learning, and holographic or hyperdimensional memory (Ororbia and Kelly 2022). In robotics, predictive frameworks encourage multimodal world models and closed perception-action loops rather than a sequence of independent classifiers and controllers. The same generative logic has been used to recast global-workspace accounts of conscious access as hierarchical partially observable Markov decision processes under active inference, giving perception-level broadcasting an explicit inferential form (Whyte and Smith 2021).

Predictive processing is nevertheless a family of related claims, not a single algorithm. Predictive coding, the free-energy principle, and active inference differ in explanatory level and should not be used interchangeably (Friston et al. 2017; Pezzulo et al. 2022). A useful architecture-level test is intervention: changing predicted sensory consequences, precision, or preference structure should produce systematic changes in perception, information seeking, and action (Proietti et al. 2025). Without such process-level evidence, predictive terminology remains an interpretive gloss rather than an implemented mechanism.

2.5 Attention Schema Theory: From Selection to Self-Modelled Control

Attention and an attention schema are theoretically distinct. Attention is a resource-allocation process that amplifies some information and suppresses other information. Attention Schema Theory (AST) proposes an

additional, simplified internal model describing that process - what the agent is attending to, how strongly, and with what control possibilities (Graziano and Webb 2015). The analogy is to a body schema: a model need not reproduce every physical detail to improve control. This distinction is important for AI because transformer attention weights or a salience map do not, by themselves, constitute an attention schema.

Wilterson and Graziano (2021) provided a direct computational test. A reinforcement-learning agent with an explicit schema of its visuospatial attention learned control more effectively and performed better after training than an otherwise comparable agent without the schema. The theory further predicts that such a model could support coordination with, and inference about, the attention of other agents (Graziano and Webb 2015). In active-inference terms, attention is often related to precision weighting, whereas the attention schema is a higher-order model that can predict and regulate that weighting.

AST therefore generates a strong implementation criterion: the system must contain a representation of its own attentional state that can be accessed by control or social-inference processes, not merely an attention mechanism (Graziano and Webb 2015). The decisive experiment compares matched agents with and without the schema under distraction, partial observability, competing goals, and multi-agent coordination (Wilterson and Graziano 2021). Benefits should appear in adaptive allocation, explanation, transfer, or recovery - not only in task accuracy on the training distribution.

2.6 Curiosity-Driven Exploration and Intrinsic Motivation

Curiosity provides an architecture with reasons to acquire information when external rewards are absent, sparse, or deceptive. Computational accounts distinguish several non-equivalent intrinsic signals: novelty, prediction error or surprise, expected information gain, learning progress, competence improvement, and empowerment or controllability (Modirshanechi et al. 2023; Oudeyer and Kaplan 2007; Ten et al. 2022). The distinction matters because each signal produces a different exploration policy. Raw prediction error may trap an agent in stochastic environments, novelty may reward unproductive distraction, and information gain may be computationally expensive.

Within a cognitive architecture, intrinsic motivation should be separated from the learning algorithm it modulates. A curiosity process evaluates candidate experiences; an action-selection process trades intrinsic and extrinsic values; memory determines whether the experience is genuinely new; a world model estimates uncertainty or progress; and metacognition regulates resource use and risk. Nagashima et al. (2024) show that pattern-discovery curiosity in ACT-R can improve or impair task completion depending on the agent's level of deliberation, illustrating that curiosity is not uniformly beneficial.

For lifelong and open-ended agents, the theoretically important quantity is productive epistemic change: improvement in a reusable model or skill, not simply unusual input (Oudeyer and Kaplan 2007; Ten et al. 2022). A robust design should therefore track learning progress, cap exploration in unsafe regions, distinguish reducible from irreducible uncertainty (Modirshanechi et al. 2023), and permit task goals or human constraints to veto curiosity. Evaluation should measure coverage, learning efficiency, transfer, and safety under sparse reward, together with failure modes such as novelty loops and catastrophic interference.

2.7 Episodic-Memory-Augmented Agents and Complementary Learning Systems

Memory is not a single store. Complementary Learning Systems theory explains why adaptive cognition benefits from at least two learning regimes: a fast hippocampal-like system that captures specific episodes after one or few exposures, and a slower neocortical-like system that extracts stable regularities across experiences (McClelland et al. 1995). Cognitive architectures add working memory for current control, procedural memory for skills, and explicit retrieval mechanisms. This differentiation is especially important for LLM agents, whose context window

is transient and whose parameters are neither a transparent autobiographical record nor an easily corrected semantic store.

An episodic memory must preserve more than semantically similar text. Architecture-level desiderata include temporal and situational binding, a record of the agent's own actions and outcomes, encoding specificity, cue-dependent and deliberately constructed retrieval, source provenance, and the possibility of consolidation, revision, or forgetting. Fountas et al. (2025) argue that episodic memory is the missing component for long-term LLM agents, while Wray et al. (2025) show that popular memory streams often retrieve by generic embedding similarity and therefore fail important cognitive requirements. A systematic survey of LLM-agent memory mechanisms reaches a compatible conclusion from the engineering side: it organizes implementations by memory sources, forms, and operations and identifies evaluation, consolidation, and management of long-term memory as open problems (Zhang et al. 2025).

Current systems explore event segmentation by surprise, hippocampal-style indexing, deep episodic representations, autobiographical narrative, and links between episodic retrieval and global-workspace control (Bärmann et al. 2021; Fountas et al. 2025; Huang et al. 2024; Peller-Konrad et al. 2023; Prescott and Dominey 2024). The unifying architecture is a cycle of event boundary detection, one-shot encoding, cue construction, retrieval, evaluation, consolidation into semantic or procedural knowledge, and controlled forgetting. Provenance and contradiction handling are indispensable because an agent must be able to distinguish remembered evidence from generated reconstruction.

Discriminating evaluation should include delayed and interference-sensitive recall, temporal-order and source-binding tasks, correction after false information, long-horizon planning that requires past episodes, and ablations of consolidation and forgetting. A vector database that improves retrieval may be useful, but it should not be described as episodic memory unless it satisfies explicit temporal, agent-centered, and reconstructive criteria (Prescott and Dominey 2024).

2.8 Dual-Process Reasoning and Planning under Metacognitive Arbitration

Within the reasoning channel, four inference types can be separated and mapped to architectural mechanisms. Deductive, monotonic inference is well served by production-rule systems such as Soar, complemented by NARS, whose non-axiomatic logic reasons under insufficient knowledge and bounded resources. Abductive, best-explanation inference draws on CLARION dual implicit-explicit processing and on Sigma graphical mixed symbolic-probabilistic messages. Analogical transfer maps onto ACT-R declarative retrieval and spreading activation together with LIDA global-workspace broadcast, while inductive generalization builds on the LIDA learning cycle and on the CoALA formulation of language agents as memory-plus-control systems. Belief-desire-intention (BDI) architectures supply the deliberative commitment layer for orchestration, and Graph-RAG provides the neuro-symbolic context interface that grounds inference in an external knowledge base. These commitments, with their governance roles, are consolidated in Section 5.5. These architectural groundings are supported by recent cognitive-systems work: metaknowledge elicitation in large language models (Longo et al. 2025), the VSM-ACTR 2 metacognitive decision model (Wu, Oltramari, et al. 2025), neuro-symbolic cognitive-consistency verification (Jo and Kwon 2025), cognitive neurosymbolic AI for decision-making (Zi et al. 2025), and human-cognition-inspired knowledge-graph retrieval (Cheng et al. 2026).

Dual-process theories distinguish fast, associative, low-cost processing from slower, capacity-demanding, explicitly controlled cognition. Contemporary theory treats these properties as continua and interacting modes rather than two anatomically or computationally isolated boxes (Evans and Stanovich 2013). Cognitive architectures are well suited to make the distinction precise because they can assign representations, algorithms, and timing to each mode. CLARION historically combined implicit and explicit processes; current hybrid systems

frequently cast a neural policy or LLM as the fast proposer and a symbolic reasoner, search procedure, or formal planner as the slow evaluator (Conway-Smith and West 2022; Sun 2024).

Planning makes the value and limitations of this pattern visible. SOFAI coordinates fast and slow planning through a metacognitive layer, and MAP decomposes prefrontal-like planning functions - conflict monitoring, state prediction, evaluation, task decomposition, and coordination - rather than expecting one LLM call to integrate them reliably (Fabiano et al. 2024; Webb et al. 2025). MERLIN2 similarly shows a practical trade-off: an LLM planner can improve interactional flexibility while a classical planner preserves plan quality and constraint satisfaction (González-Santamarta, Rodríguez-Lera, Fernández-Llamas, et al. 2024).

The theoretically decisive component is arbitration. A dual-process architecture should specify when and why it switches, escalates, combines outputs, or abstains. Relevant variables include uncertainty, novelty, expected loss, reversibility, time pressure, computational cost, and the expected value of additional reasoning. A complementary formalization casts cognitive control itself as precision optimization within active inference, in which a single control signal balances deliberative and habitual contributions to action selection (Proietti et al. 2025). Metacognition in the Nelson-Narens tradition separates an object level that performs cognition from a meta level that monitors and controls it (Nelson and Narens 1990). The CMC extension and a recent systematic review of metacognitive architectures translate this distinction into explicit self-representations, memory for the agent's own thoughts, and control operations (Laird, Lebiere, Rosenbloom, and Stocco 2026; Ukov and Tsochev 2025).

Evaluation must therefore compare a dual-process system not only with an LLM-only or symbolic-only baseline, but also with a matched hybrid lacking calibrated arbitration. Key outcomes include selective accuracy, constraint violations, planning horizon, recovery, latency, and resource use. If the controller cannot predict when slow cognition adds value, the extra architecture may increase complexity without improving reliability (Fabiano et al. 2024).

2.9 Computational Theory of Mind and Bounded Social Inference

Theory of mind (ToM) concerns the ability to explain and predict behavior by representing mental states such as beliefs, desires, intentions, knowledge, and uncertainty (Premack and Woodruff 1978). Early artificial systems often encoded social rules or fixed belief-desire-intention structures. Contemporary computational ToM increasingly treats social understanding as inverse planning: the observer infers latent goals, beliefs, costs, or competence from another agent's actions under a probabilistic generative model (Jara-Ettinger 2019).

This Bayesian account aligns social cognition with the broader architecture of generative inference but adds distinctive representational requirements. The system must distinguish self from other, maintain agent-indexed beliefs, represent false or incomplete beliefs, update models from behavior, and preserve uncertainty over competing explanations. Z. Wang, Davis, et al. (2025) emphasize abstraction and development: tractable social cognition relies on reusable primitives and simplified context-specific models rather than unlimited recursive mindreading. Erdoğan et al. (2025) operationalize this as computational theory of mind with abstraction, in which higher-level social constructs such as norms and roles are abstractions of beliefs that make recursive reasoning tractable in human-agent collaboration. Lebiere, Pirolli, et al. (2024) demonstrate a complementary cognitive-model approach using ACT-R mechanisms and model tracing to personalize predictions of another agent.

ToM also connects attention, communication, and memory. An attention schema may support estimates of what another agent can see or prioritize (Graziano and Webb 2015); episodic memory supplies interaction history; and pragmatic language understanding requires beliefs about shared knowledge and intent. For practitioners, however, social inference creates privacy, stereotyping, manipulation, and over-attribution risks. A bounded ToM module should expose its evidence, uncertainty, scope, and permitted uses, and it should avoid inferring sensitive traits when task performance does not require them.

Evaluation should combine controlled false-belief or inverse-planning tasks with ecologically valid interaction, held-out agents, adversarial behavior, and calibration of social predictions. The key comparison is not whether an LLM can produce plausible mental-state language, but whether an explicit, revisable social model improves prediction and collaboration without increasing harmful inference.

2.10 Ability-by-Ability Framework Synthesis

Sections 2.3–2.9 examined the main theoretical traditions; this subsection condenses what they imply across the twelve channels. Table 2 states, for each channel, the theoretical anchor, the architecture-level commitment, and the evidence that would distinguish it from a simpler alternative, and Tables S1–S3 list the frameworks and implementing systems. Four cross-channel regularities stand out.

First, perception, attention, and action selection are increasingly specified within one generative, precision-weighted loop rather than as a feed-forward pipeline (Ciria et al. 2021; Friston et al. 2017; Pezzulo et al. 2022; Proietti et al. 2025; Whyte and Smith 2021). Systems such as CogNGen and CORTEX realize parts of this loop (Ororbia and Kelly 2022; Trinidad Barnech et al. 2024), and Attention Schema Theory adds an accessible model of the agent's own attention (Graziano and Webb 2015; Wilterson and Graziano 2021).

Second, memory and learning converge on complementary learning systems: fast episodic capture, slower consolidation, deliberate cue-based retrieval, and controlled forgetting (Fountas et al. 2025; McClelland et al. 1995; Pecora et al. 2026; Wray et al. 2025; Zhang et al. 2025; Zhou, Z. Liu, et al. 2024). Curiosity and learning progress act as intrinsic drivers whose effect depends on the agent's level of deliberation (Modirshanechi et al. 2023; Nagashima et al. 2024; Ten et al. 2022), and knowledge supplied by language models enters only through verification (Kirk et al. 2024; Laird, Wray, et al. 2024; Ororbia and Kelly 2024).

Third, reasoning, planning, and action selection share a dual-process organization in which a fast neural or language-model proposer is checked by a slower symbolic or model-based process under an arbitrating controller (Fabiano et al. 2024; González-Santamarta, Rodríguez-Lera, Fernández-Llamas, et al. 2024; Sun 2024; Webb et al. 2025; West et al. 2024; Zhu and Simmons 2024). Metacognition, specified as reasoning over explicit representations of the agent's own capabilities and processes, supplies that controller (Laird, Lebiere, Rosenbloom, and Stocco 2026; T. Liu and van der Schaar 2025; Nelson and Narens 1990; Ukov and Tsochev 2025) and is now operationalized in planning and code-generation agents (Jian et al. 2025; Zhou, Y. Liu, et al. 2025).

Fourth, communication and social cognition are bounded rather than open-ended. Language enters the cognitive cycle through a language-model front end that must be grounded and verified (Allgeuer et al. 2024; Scheibl et al. 2025; Sievers and Russwinkel 2025; Sun 2024), and other agents are represented by simplified, context-specific, uncertainty-aware models rather than unlimited recursive mindreading (Erdoğan et al. 2025; Lebiere, Pirolli, et al. 2024; Z. Wang, Davis, et al. 2025). Neuro-symbolic integration is the seam that connects all of these channels; its open questions are where the seam lives and what must cross it (Acharya and Song 2025; DiGilio and Agrawal 2026; Nadimuthu et al. 2026; Safron 2020; Wan et al. 2025; W. Wang and Yang 2024), which motivates the interface commitments developed in Section 5.

2.11 An Integrative Post-Kotseruba Framework: Governed Hybrid Agency

The theoretical traditions above can be integrated without claiming that one mechanism explains all cognition. Their common implication is architectural: intelligent behavior requires coordinated control across timescales, representations, and sources of uncertainty. A post-2018 cognitive architecture can be represented as five interacting layers: a generative sensorimotor interface (Friston et al. 2017); complementary working, episodic, semantic, and procedural memories (McClelland et al. 1995); fast proposal and slow verification pathways (Evans and Stanovich 2013); a metacognitive governance layer that allocates attention, computation, trust, and exploration (Nelson and Narens 1990); and a social-normative layer that represents other agents, roles, consent,

and constraints (Jara-Ettinger 2019). Neuro-symbolic interfaces connect the layers and must carry semantic type, provenance, uncertainty, and intervention rules (Acharya and Song 2025).

Table 2. Integrative theoretical framework linking cognitive abilities to architectural commitments and discriminating evidence.

Ability channel	**Primary theoretical anchor(s)**	**Architecture-level commitment**	**Minimum discriminating evidence**
Perception	Predictive processing; active inference; embodied affordances	A multimodal generative world model that exposes uncertainty and links observations to possible actions	Calibration and counterfactual state prediction under occlusion and distribution shift; generative-model ablation
Attention	Attention Schema Theory; precision weighting	A selection mechanism plus an accessible internal model of what is prioritized and why	Schema-versus-attention-only ablation under distraction, partial observability, and social coordination
Action selection	Expected free energy; utility/operator selection; dual-process arbitration	Layered reactive or habitual policies with deliberative override, clarification, and recovery	Adaptive switching under novelty, risk, and time pressure; constraint violations and recovery
Memory	Complementary Learning Systems; episodic reconstruction; spreading activation	Working, episodic, semantic, and procedural stores with cue-based retrieval, provenance, consolidation, and forgetting	Delayed recall, temporal/source binding, interference, correction, and long-horizon task dependence
Learning	Curiosity and intrinsic motivation; Hebbian/free-energy learning; lifelong learning	Extrinsic and intrinsic value, world-model update, semantic/procedural consolidation, and forgetting control	Sparse-reward exploration, learning progress, transfer, retention, and resistance to novelty traps
Reasoning	Dual-process and neuro-symbolic inference; analogy; non-monotonic constraints	Fast candidate generation with explicit compositional representations, verification, and revision	Out-of-distribution compositional generalization, contradiction detection, intervention, and verifier ablation
Planning	Hierarchical/factored planning; model-based search; dual-process control	Task decomposition, state prediction, constraint checking, execution monitoring, and replanning	Long-horizon success, state-validity and safety constraints, latency/cost trade-offs, and disturbance recovery
Metacognition	Nelson-Narens monitoring/control; CMC extension; hierarchical active inference	Explicit representations of capabilities, confidence, strategies, and outcomes that can alter lower-level processing	Calibration, selective prediction, self-repair, strategy switching, and causal intervention on the meta level

Ability channel	Primary theoretical anchor(s)	Architecture-level commitment	Minimum discriminating evidence
Communication	Grounded language; pragmatic inference; LLM as implicit competence	Language bound to current world state, memory provenance, shared knowledge, and executable schemas	Pragmatic generalization, clarification, grounding, provenance, and safe mapping from utterance to action
Motivation/emotion	Curiosity, learning progress, empowerment, expected value of control	Control signals that allocate exploration, effort, persistence, and affective modulation across modules	Productive exploration versus distraction; goal conflict, risk sensitivity, and persistence under sparse reward
Social cognition	Bayesian ToM/inverse planning; BDI; abstraction	Bounded, agent-indexed, uncertainty-aware models of beliefs, goals, knowledge, and roles	False-belief/inverse-planning tasks, held-out agents, calibration, privacy constraints, and social-model ablation
Neuro-symbolic integration	CMC bridging; typed hybrid interfaces; probabilistic-symbolic coupling	Bidirectional exchange between learned and explicit representations with provenance, uncertainty, and veto paths	Uncertainty propagation, consistency, compositional transfer, intervention success, and directional-coupling ablations

Proposition 1 - Architecture is cross-temporal coordination. The distinctive contribution of a cognitive architecture is not maximal component count but the maintenance of coherent state, learning, and control across episodes, tasks, and disturbances (Newell 1990).

Proposition 2 - Foundation models are competence reservoirs, not complete architectures. They provide broad conditional generation but require external commitments about persistent memory, belief acceptance, goals, action authorization, and learning (Wray et al. 2025).

Proposition 3 - Memory is the binding mechanism of agency. Episodic records connect decisions with consequences, semantic consolidation supports generalization, and procedural memory preserves acquired skills. Without source-sensitive memory, metacognition and accountability are shallow (McClelland et al. 1995).

Proposition 4 - Metacognitive arbitration turns hybridization into governance. Combining neural and symbolic components is insufficient unless a process monitors uncertainty, risk, and resource cost and can alter which component controls behavior (Fabiano et al. 2024).

Proposition 5 - Attention and motivation jointly allocate bounded resources. Attention determines what is processed now; intrinsic and extrinsic motivation determine which information, skills, and goals deserve future processing. Their coupling should be explicit and testable (Graziano and Webb 2015; Oudeyer and Kaplan 2007).

Proposition 6 - Social cognition must remain bounded and uncertainty-aware. Useful collaboration requires models of other agents, but safe implementation requires scope limits, evidence, calibration, privacy protection, and the ability to revise or abstain (Z. Wang, Davis, et al. 2025).

Proposition 7 - The neuro-symbolic seam is an interface problem. Reliable hybrid agency depends on typed messages that distinguish observations, hypotheses, memories, beliefs, candidate plans, approved plans, and actions, together with provenance and uncertainty (Acharya and Song 2025).

Proposition 8 - Ability claims require interaction-sensitive evaluation. An architecture should be tested not only on isolated memory, planning, or ToM tasks, but on whether the proposed mechanism improves the complete perception-memory-reasoning-action loop under intervention and ablation.

Sections 2.3–2.10 examine each theoretical tradition and each ability channel in turn. Two synthesis tables make the resulting mapping explicit and provide the ability-by-ability comparison targets promised in Section 1.3. Table S1 records, for every ability channel, the framework that most strongly shaped post-2020 architectural work, the competing or complementary accounts that remain live, and representative systems that instantiate the dominant commitment. Table S2 inverts the axis: it lists each framework together with its core computational commitment, the architectures that implement it, and the ability channels it reaches. The two views together show a field converging on a shared decomposition of problems while retaining theoretically meaningful competition over mechanisms.

Note. The two tables synthesize the post-2018 corpus and the framework-level literature reviewed in Sections 2.3–2.10; a system is listed under a framework when its published description contains an explicit mechanism realizing that framework's core commitment, not merely the framework's vocabulary.

2.12 Boundary Conditions, Cognitive Fidelity, and Falsifiability

The framework is functional and architectural; it does not imply that an artificial system has human consciousness, emotion, selfhood, or social understanding merely because it contains similarly named mechanisms. Cognitive inspiration concerns the transfer of organizational principles. Cognitive fidelity is a stronger empirical claim requiring correspondence with human behavior, development, neural data, or process measures (Hake et al. 2022). Engineering utility is different again: a mechanism may improve reliability without accurately modeling human cognition. These claims should be stated separately.

The framework is also falsifiable at several levels. A proposed ability can fail the representation test because no relevant state is maintained; the process test because the state is never used causally; the coordination test because the mechanism is bypassed in the full system; or the evidence test because a simpler baseline performs equally well. This layered standard prevents cognitive terminology from becoming decorative and allows classical architectures, foundation-model agents, and neuro-symbolic systems to be compared on common causal questions.

3 Methods

This section describes the focused, reproducible search, screening, and coding procedures, together with the reliability checks and the disclosed use of AI tools in the review workflow.

3.1 Review Design and Reporting Approach

This study is a critical integrative review (Whittemore and Knafl 2005) supported by a focused, reproducible search and evidence-mapping procedure over two citation databases — Scopus and the Web of Science Core Collection. PRISMA 2020 (Page et al. 2021) is used to report information sources, record flow, eligibility decisions, and exclusions transparently; it is not invoked to imply that the review exhaustively covers all post-2018 cognitive-architecture research. The focused database corpus is designed primarily to answer questions about the integration and governance of foundation models within architecture-level cognitive systems, and the query therefore required a foundation- or generative-model term. Foundational cognitive-architecture theory and relevant post-2018 work outside the focused query are treated as a separately identified contextual synthesis rather than as part of the counted corpus. These databases were chosen for their complementary, quality-controlled indexing of the artificial-intelligence, software-engineering, and cognitive-science literatures; preprint-only servers were deliberately excluded so that every included record carries indexed bibliographic metadata. The protocol was

designed to be reproducible: the full Boolean strings, field tags, and filters are reported below, and record-level screening decisions are available as supplementary material.

3.2 Information Sources and Search Strategy

Eight thematic strands informed the construction of the three search blocks:

"cognitive architecture core abilities post-2018 review"
"Soar ACT-R CLARION LEABRA developments 2019 2020 2021 2022 2023 2024"
"cognitive architecture large language model integration"
"neuro-symbolic cognitive architecture reasoning"
"cognitive architecture robotics perception attention memory"
"cognitive architecture metacognition planning motivation"
"predictive processing active inference cognitive architecture"
"theory of mind social cognition computational architecture"

Scopus and the Web of Science Core Collection were both searched on 10 July 2026, and the result sets were exported and frozen on the same day to fix the record counts against subsequent database updates. In Scopus the query was issued over the TITLE-ABS-KEY field, and in Web of Science over the Topic (TS) field. Three concept blocks were combined with AND: cognitive architectures and cognitive systems (Block A); foundation and generative models — large language, vision-language, and generative models (Block B); and governance-relevant mechanisms — agency, reasoning, planning, metacognition, memory, grounding, verification, and autonomy (Block C). Results were restricted to English-language journal articles, reviews, and conference papers published from 2018 onward. The complete query strings are reported in Appendix A.

3.3 Eligibility Criteria and Selection

Records were eligible when they (a) addressed a cognitive architecture, cognitive system, or architecture-level agentic framework; (b) substantively engaged foundation or generative models as a component, comparison, or critique; and (c) discussed at least one governance mechanism — persistent memory, metacognitive control, planning, verification, perceptual grounding, normative or social modelling, or neuro-symbolic integration. Benchmark-only papers without an architectural claim, editorials and position pieces without new claims, posters and abstracts shorter than four pages, work using a cognitive architecture only as an unexplained black box, and non-English records were excluded.

The two searches returned 941 records from Scopus and 232 from Web of Science, giving 1,173 records in total. After removal of 143 duplicates, 1,030 unique records were screened. A reproducible, rule-based pre-filter retained the 660 records whose title or abstract combined a cognitive-architecture or agent-framework term with a foundation-model term, excluding 370 records (proceedings front matter, out-of-scope keyword matches, and records without foundation-model engagement). The 660 retained records were then assessed at the title-and-abstract level against the three inclusion criteria by a single-pass, LLM-assisted eligibility assessment that was run independently of the pre-filter and is specified in full in Section 3.6 and Table 3; 541 records were excluded as domain applications, general language-model or agent methods, or conceptual essays without a cognitive-architecture contribution, yielding 119 included works. The first author did not re-screen all 660 records individually; instead, the assessment was audited against author judgement on a purposive borderline subset and checked for run-to-run reproducibility on a random sample (Section 3.6). Full texts of the 119 included works were retrieved and consulted during data extraction and synthesis, which led to no further exclusions (Figure 1). Accordingly, the 119-record set is treated as a machine-assisted focused evidence map rather than an exhaustively human-screened systematic corpus: because not all 660 candidate records were independently human-rescreened and the borderline audit identified both potential false exclusions and false inclusions, the corpus frequencies

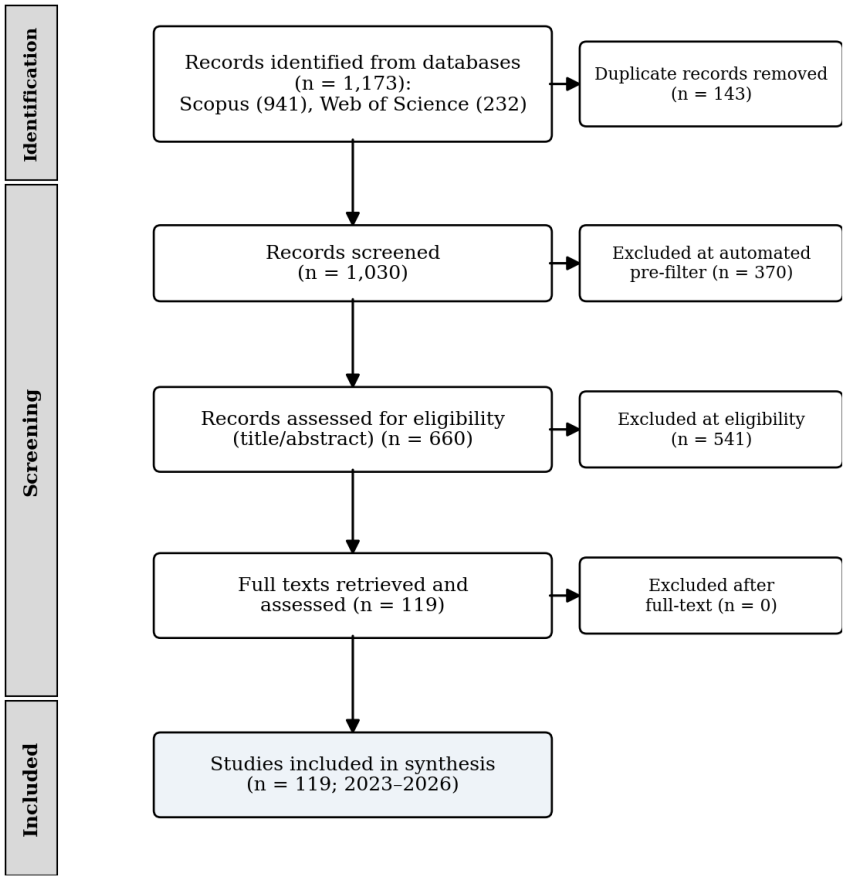


Fig. 1. PRISMA 2020 flow for the Scopus and Web of Science search: from 1,173 identified records (941 Scopus, 232 Web of Science), through duplicate removal, an automated pre-filter, and content-based eligibility screening, to the 119 works included in the synthesis.

and coverage claims reported below are descriptive and are not interpreted as unbiased estimates of the entire post-2018 literature.

The synthesis draws on two distinct layers. The core focused corpus comprises the 119 works meeting all three inclusion criteria and grounds the quantitative evidence map. A contextual layer of foundational and theoretical works — including pre-2018 sources and post-2018 theory whose abstracts do not use foundation-model terms — was used to reconstruct the twelve ability-channel frameworks but is not counted in the 119-work corpus. Quantitative statements refer to the core corpus unless attributed to a specific cited source.

Two verification steps were applied. First, the numerical flow was re-audited for arithmetic consistency (941 + 232 = 1,173 identified; 1,173 − 143 = 1,030 screened; 1,030 − 370 = 660 title/abstract-screened; 660 − 541 = 119 synthesized). Second, reference metadata for all cited works were re-checked against publisher records in July 2026; any citation first circulating as a preprint was updated to its indexed, peer-reviewed version where available, or removed if no indexed venue could be confirmed.

3.4 Data Extraction and Coding

For each retrieved full text, structured extraction captured the architecture or system name; publication status; addressed ability channels; theoretical framework; role assigned to any LLM or foundation model; evaluation setting; application domain; relationship to a classical lineage; and source-supported statements relevant to benefits, limitations, and research gaps. The full texts of the included works, rather than model-generated summaries, were treated as the evidential basis for the synthesis.

The twelve-channel coding frame extended the original core-ability map to include communication, motivation/emotion, social cognition, and neuro-symbolic integration as explicit analytic dimensions. Foundation-model integration was coded inductively and consolidated into five patterns: (1) knowledge or generative memory, (2) language and multimodal interface, (3) heuristic proposal or planning, (4) rule/model bootstrapping, and (5) bidirectional neuro-symbolic coupling. Applications were grouped into robotics and human-robot interaction, manufacturing and industrial collaboration, cyber-physical systems, healthcare and assistive systems, cyberdefense, recommender/personalization systems, and cognitive simulation/education. Governance-mechanism mentions were additionally coded at the title-and-abstract level across the 119 included works; these abstract-level frequencies (reported in Section 4.1 and the supplementary evidence table) indicate that metacognitive control is the least explicitly foregrounded mechanism, appearing in 11 of 119 abstracts (9%), whereas reasoning (70), memory (51), and grounding (38) appear far more often. Because the coding reflects abstract-level salience rather than verified architectural coverage, these counts indicate emphasis rather than exhaustive presence.

3.5 Evidence Maturity and Synthesis

Because the evidence spans conceptual proposals, formal models, simulations, robot demonstrations, and heterogeneous comparative experiments, a conventional risk-of-bias tool or meta-analysis would be inappropriate. Instead, implementation maturity was described using four non-hierarchical categories: specification (conceptual or formal architecture), prototype (implemented primarily in simulation or a constrained task), demonstration (physical or operational deployment without a strong comparator), and comparative evaluation (baseline comparison, ablation, or benchmarked performance). These categories indicate the kind of evidence available, not study quality. Synthesis proceeded in six passes: historical-theoretical reconstruction; ability-channel update; framework-to-architecture mapping; LLM-integration pattern analysis; application and evaluation mapping; and abductive construction of the Architectural Governance Stack from recurring cross-channel requirements.

3.6 LLM-Assisted Eligibility Assessment and Other Uses of Generative AI

The record flow separated a deterministic component from a model-assisted one. The pre-filter was a reproducible, rule-based procedure with no generative-AI involvement. A large language model then performed the title-and-abstract eligibility assessment of the 660 pre-filtered records and assisted the extraction of structured metadata; it was also used, under author direction, for language editing. The first author defined the eligibility criteria and the rubric, audited the assessment on a calibration subset and a validation sample, consulted the full texts of all included records, and performed the final synthesis with the second author. Table 3 summarizes the protocol, which is specified in the remainder of this section.

Model, interface, and settings. The assessment used Claude Sonnet 5 (Anthropic), called programmatically through the Claude Agent SDK in July 2026. Decoding parameters (temperature, nucleus sampling, and maximum output length) were not overridden and therefore took the SDK defaults. The exact model version string, the SDK version, and the resolved parameter values were not logged, which limits exact technical replication of the configuration. We report this explicitly because default settings do not guarantee deterministic output; run-to-run reproducibility was therefore measured empirically rather than assumed (see below).

Input, rubric, and decision rule. Each record was presented to the model as its title and abstract, together with a fixed rubric that operationalizes the three inclusion criteria of Section 3.3 as conjunctive conditions: a record was to be included only if it was substantively about a cognitive architecture, cognitive system, or architecture-level agent framework; substantively engaged a foundation or generative model as a component, comparison, or critique; and addressed at least one governance-relevant mechanism. The rubric instructs the model to apply these conditions strictly and lists six exclusion categories: proceedings front matter; domain applications in which an LLM is used only as a tool; general LLM or agent method and benchmark papers without a cognitive-systems framing; purely conceptual essays without a concrete architecture-plus-foundation-model integration; explainability or knowledge-representation papers without a cognitive architecture; and false keyword matches (for example, *soar* used as a verb or *sigma* as a statistic). The model returned a structured INCLUDE or EXCLUDE label with a three-to-six-word rationale, and records it judged ambiguous defaulted to EXCLUDE. Each record was assessed exactly once, and that single label was the eligibility decision: no majority voting, self-consistency sampling, or confidence threshold was applied. Every decision and rationale is logged under a stable record identifier that links each record from identification through final coding.

Agreement with author judgement. The first author independently screened a purposive 30-record subset, deliberately enriched for borderline cases and drawn from both sides of the rule-based pre-filter; the author would have included 5 of the 30 records. Nineteen of the 30 records had passed the pre-filter and were therefore assessed by the LLM in the main run, including all five records the author would have included; the pre-filter thus removed none of the author's inclusions in this subset. On these 19 records the LLM assessment agreed with the author on 12, all of them exclusions: it excluded all five records the author would have retained and retained two the author would have excluded. The remaining 11 records had been removed by the pre-filter and were also excluded by the author; they bear on the pre-filter but not on the LLM assessment, and are reported separately in the calibration file. The calibration was used to characterize the assessment, not to correct it: the five records the model excluded were not added to the corpus and the two it retained were not removed, so that the corpus remains the output of a single, fixed protocol; all seven are identified in the calibration file. Because the subset is purposive and contains few eligible records, these counts are not corpus-wide estimates of accuracy. What the subset does show is the direction of error: on borderline records the assessment errs towards exclusion, as expected from its strict rubric and its ambiguous-to-exclude rule. The corpus is therefore more likely to under-represent borderline eligible work than to contain ineligible work, and we treat its coverage accordingly (Section 8).

Run-to-run reproducibility. A second pass with the same rubric, run separately from the first, was applied to a random sample of 50 of the 660 pre-filtered records. The two passes agreed on 49 of 50 records (98%; Cohen's $\kappa = 0.90$): five records were included by both passes, 44 were excluded by both, and one was included by the second pass only. This check measures the consistency of the automated assessment across independent runs, not its agreement with an external human standard; the two checks are therefore complementary rather than interchangeable.

Other uses and responsibility. Model-assisted metadata extraction was verified against the source records, and the full texts of the included works, rather than model-generated summaries, were treated as the evidential basis for the synthesis (Section 3.4). The rubric, model settings, the 660 logged decisions, the 30-record calibration table, and the 50-record validation set are released in the reproducibility package listed under Data availability.

4 Results

This section reports the evidence map: the ability-by-ability synthesis, the profile of the 119-work corpus, the coverage matrix, and the recurring foundation-model integration patterns.

Table 3. LLM-assisted eligibility assessment: protocol summary.

Element	Specification
Records assessed	660 records retained by the rule-based pre-filter (title and abstract)
Model and access	Claude Sonnet 5 (Anthropic), Claude Agent SDK, July 2026
Decoding settings	SDK defaults; temperature, nucleus sampling, and output length not overridden
Rubric	Three conjunctive inclusion conditions and six exclusion categories (Section 3.6); full text in the reproducibility package
Output	INCLUDE or EXCLUDE with a three-to-six-word rationale
Decision rule	Single pass per record; ambiguous records default to EXCLUDE; no voting or thresholding
Result	119 included, 541 excluded
Author calibration	Purposive, borderline-enriched subset ($n = 30$); the 19 records that passed the pre-filter: agreement 12/19, all on exclusions; 5 false exclusions, 2 false inclusions; pre-filter removed none of the 5 author inclusions
Reproducibility	Random sample ($n = 50$), two independent passes: agreement 49/50, $\kappa = 0.90$
Released artefacts	Rubric, settings, per-record decisions with rationales, calibration and validation tables

The counts reported in this section are bounded by the review design. Because the query required a foundation- or generative-model term, the corpus characterizes foundation-model-related cognitive-architecture work, not the prevalence of mechanisms across all post-2018 cognitive-architecture research. Because eligibility was determined by an LLM-assisted assessment that was not followed by independent human rescreening of every candidate record, the frequencies below describe the retained evidence map and should be read as indicative patterns rather than as unbiased population estimates. Where a percentage is reported, it refers to the 119 retained works unless a specific external source is cited.

4.1 Corpus Profile and Five Field-Level Shifts

The 119-work evidence map contains extensions of classical architectures, newly proposed architectures, focused mechanism studies, reviews, and architecture-level integrations. Many included works treat foundation models as first-class architectural components, whereas only eleven foreground metacognition in their title or abstract — 9% of the corpus, consistent with the abstract-level frequencies reported in Section 3.4 and the supplementary evidence table. All included works are indexed in Scopus or Web of Science; no implemented system demonstrated all twelve ability channels at comparable maturity. By publication type, 36 of the 119 works (30%) are journal articles and 83 (70%) are conference or workshop papers, distributed across 2023 (10), 2024 (15), 2025 (47), and 2026 (47). Abstract-level mechanism coverage within the corpus is led by reasoning (70; 59%), memory (51; 43%), grounding (38; 32%), and social or normative modelling (31; 26%), followed by verification (25; 21%), planning (19; 16%), neuro-symbolic integration (16; 13%), lifelong learning (15; 13%), and metacognition (11; 9%). By implementation maturity, the corpus comprises 47 comparative evaluations (39%), 27 conceptual or formal specifications (23%), 25 system demonstrations (21%), and 20 simulation prototypes (17%). Foundation models are integrated chiefly as heuristic proposers or planners (39; 33%), as generative or knowledge memory (24; 20%), through bidirectional neuro-symbolic coupling (22; 18%), and as language or multimodal interfaces (21; 18%), with rule- or model-bootstrapping less common (7; 6%) and six works (5%) using a foundation model only as a comparator. For this profile each work was assigned a single primary integration pattern, although more advanced systems could instantiate additional secondary patterns. The full profile and per-record coding, including the multi-label abstract-level mechanism flags for all 119 works, are provided as supplementary material (`corpus119_full_coding.csv`). This multi-label distribution is the quantitative component of the answer to RQ2:

because it is derived from titles and abstracts rather than verified full-text coverage, the twelve-channel account should be read as an abstract-level frequency map complemented by the qualitative synthesis of Sections 2.3–2.10 and the representative-architecture coverage matrix (Table 5), not as a validated full-text census of every channel in every system.

Five field-level changes organized the corpus:

Hybridization became recurrent across the corpus and dominant in the reviewed autonomous-robotics subset. Symbolic planning, production rules, and explicit memories are increasingly combined with deep perception, learned policies, vector representations, LLMs, and vision-language models. In the focused robotics review, which screened 502 records and accepted 22 studies, 13 of the 22 accepted studies (59.1%) used hybrid symbolic-subsymbolic architectures (González-Santamarta, Rodríguez-Lera, and Matellán-Olivera 2025).

Implementation infrastructure became more standardized. ROS or ROS 2 underpinned 20 of the 22 architectures (90.9%) in the same robotics sample, improving integration and code reuse while leaving higher-level ontologies, motivation, and neural-symbolic interfaces comparatively fragmented (González-Santamarta, Rodríguez-Lera, and Matellán-Olivera 2025).

LLM integration emerged as a major new design axis. LLMs now serve as knowledge sources, conversational interfaces, planners, proposal generators, and rule-authoring tools (Choi 2024; González-Santamarta, Rodríguez-Lera, Fernández-Llamas, et al. 2024; Joshi and Ustun 2024; Kirk et al. 2024).

Metacognition, episodic memory, and lifelong learning received renewed architectural attention. The CMC has been explicitly extended with metacognitive representations, several studies focus on memory for an agent's own cognition, and long-horizon agents are being redesigned around episodic capture and consolidation (Laird, Lebiere, Rosenbloom, and Stocco 2026; Ukov and Tsochev 2025); within the corpus, however, metacognition remains the least frequently foregrounded mechanism at the abstract level (9%).

Evaluation did not keep pace with architectural ambition. In the autonomous-robotics review, execution time was reported in 10 of 22 included studies (45.5%), success rate in 7 (31.8%), and explainability metrics in only 2 (9.1%), despite 18 of 22 (81.8%) targeting HRI or real-world deployment (González-Santamarta, Rodríguez-Lera, and Matellán-Olivera 2025). Metrics for transfer, retention, calibration, and adaptive reasoning remained uncommon.

4.2 Evolution of the Twelve Analytic Ability Channels

Table S3 summarizes the central change in each ability channel. The synthesis does not imply that one theory has won. Rather, it identifies the frameworks that most strongly shaped post-2018 architectural work and highlights where the implementation evidence remains narrow.

4.3 Evidence for Theoretical Convergence Without Theoretical Uniformity

The evidence map supports the CMC as the field's most stable shared vocabulary, but not as a completed implementation. Post-2018 work tests the model against neuroimaging, realizes it through generative and neuro-mimetic substrates, and extends it toward metacognitive representation (Hake et al. 2022; Laird, Lebiere, Rosenbloom, and Stocco 2026; Ororbia and Kelly 2024; Ukov and Tsochev 2025; West et al. 2024). This persistence is theoretically important: foundation models are stateless between calls, so durable memory, identity, and learning must be supplied by the surrounding architecture rather than by the model itself.

Predictive processing and active inference provide the broadest cross-ability process language, linking perception, action, attention, learning, and intrinsic information seeking through generative models, prediction errors, precision, and expected free energy (Ciria et al. 2021; Friston et al. 2017; Safron 2020). Yet implementation commitments vary considerably. Some systems instantiate local prediction-error dynamics and explicit generative models, whereas others invoke predictive terminology at a conceptual level. The evidence therefore supports a convergent design family rather than a single verified mechanism.

Dual-process organization is the most visible engineering bridge between classical architectures and foundation models. Fast neural or LLM-based proposal generation is paired with slower explicit reasoning, planning, or verification, while an additional controller is expected to allocate resources and manage escalation (Conway-Smith and West 2022; Fabiano et al. 2024; Sun 2024; Webb et al. 2025). Comparative evidence supports modular decomposition in selected planning tasks, but calibrated arbitration remains underdeveloped. Without explicit switching criteria and ablations, a dual-process label may merely redescribe heterogeneous components.

Frameworks with narrower scope often yield sharper falsifiable claims. Attention Schema Theory predicts benefits from an internal model of attention beyond attention itself (Wilterson and Graziano 2021). Curiosity frameworks distinguish novelty, information gain, learning progress, and empowerment as different drivers of exploration (Modirshanechi et al. 2023; Oudeyer and Kaplan 2007; Ten et al. 2022). Complementary-learning accounts require one-shot episodic capture and slower semantic consolidation (McClelland et al. 1995), while computational theory of mind requires explicit, revisable models of other agents rather than anthropomorphic language alone (Jara-Ettinger 2019; Lebiere, Pirolli, et al. 2024; Z. Wang, Davis, et al. 2025). Within the reviewed corpus this pattern suggests partial convergence on a common decomposition of problems, while theoretically meaningful competition over mechanisms is retained.

4.4 Five Patterns of Foundation-Model Integration

The corpus does not support a single architecture for integrating foundation models. It does reveal five recurrent patterns (Table 4), each with a distinct benefit-risk profile. The most mature systems often combine more than one pattern.

Table 4. Foundation-model integration patterns, benefits, failure modes, and governance requirements.

Pattern	Foundation-model role	Representative examples	Primary benefit	Characteristic failure mode	Minimum governance requirement
1. Knowledge or generative memory	Supplies candidate facts, semantic associations, or long-term representations	Soar knowledge acquisition; transformer integration; ACT-R retrieval support (Kirk et al. 2024; Laird, Wray, et al. 2024; Sievers and Russwinkel 2025)	Rapid access to broad prior knowledge and flexible retrieval	Confabulated or context-insensitive memories; unclear provenance	Persistent source-tagged memory, retrieval confidence, and verification before commitment
2. Language and multimodal interface	Parses instructions, conducts dialogue, or maps perception to linguistic representations	MERLIN2; grounded multimodal HRI; co-constructive task learning (Allgeuer et al. 2024; González-Santamarta, Rodríguez-Lera, Fernández-Llamas, et al. 2024; Scheibl et al. 2025)	Natural interaction and reduced domain-engineering burden	Pragmatic misunderstanding, hallucinated affordances, and unsafe action mapping	Grounding to world state, executable schema validation, clarification policy, and human override

Pattern	Foundation-model role	Representative examples	Primary benefit	Characteristic failure mode	Minimum governance requirement
3. Heuristic proposer or planner	Generates candidate plans, decompositions, or actions for a symbolic controller	SOFAI; MAP; MERLIN2; LLM-based embodied controllers (Fabiano et al. 2024; González-Santamarta, Rodríguez-Lera, Fernández-Llamas, et al. 2024; Webb et al. 2025)	Search guidance, semantic task decomposition, and flexible handling of open instructions	Invalid plans, false success claims, poor sequential compliance	Formal constraints, execution monitor, state reconciliation, fallback planner, and abstention
4. Rule, code, or model bootstrapping	Translates natural language or examples into productions, rules, or cognitive models	LLM-authored ACT-R/Soar models; bootstrapped cognitive agents (Wu, Souza, et al. 2024; Zhu and Simmons 2024)	Lower knowledge-engineering cost and faster adaptation	Incorrect rules silently enter the architecture and propagate errors	Sandboxed generation, unit tests, provenance, versioning, and reversible updates
5. Bidirectional neuro-symbolic coupling	Extracts symbolic structure from learned representations and sends symbolic constraints back to the model	CLARION-inspired coupling; CogRec-style explanation; generative-network/CMC bridges (O. J. Romero et al. 2023; West et al. 2024)	Potentially combines flexible learning with explicit reasoning and intervention	Semantic mismatch at the interface; uncalibrated uncertainty transfer	Typed interface contracts, shared ontology, calibration mapping, and intervention logging

4.5 Representative Architectures and Coverage Patterns

The coverage matrix in Table 5 is a source-based map of explicit treatment, not a performance ranking. A filled circle indicates that the cited description contains an explicit module or mechanism for the channel; an open circle indicates absence or insufficiently explicit treatment in the source. The matrix confirms two points. Classical architectures still supply broad and persistent scaffolding, while newer systems frequently specialize around a theoretical commitment or application. Conversely, broad specifications such as CMC plus metacognition cover many channels on paper but lack an implementation at equivalent maturity across them.

Table 5. Representative architecture families, proposals, and implemented systems by explicit coverage of twelve analytic ability channels.

Architecture	P	At	Act	M	L	R	Pl	MC	C	Mot	S	NS
Soar + Transformer	●	●	●	●	●	●	●	○	●	○	○	●
ACT-R + LLM cases	○	●	●	●	●	●	●	●	●	●	●	●
CLARION + LLM roles	○	○	●	●	●	●	●	●	●	●	○	●
ICARUS	●	○	●	●	●	●	●	○	○	○	○	○
CogNGen	●	●	●	●	●	○	●	○	○	●	○	○

Architecture	P	At	Act	M	L	R	Pl	MC	C	Mot	S	NS
IWMT	●	●	●	●	●	●	●	●	○	●	●	○
SOFAI	○	○	●	○	●	●	●	●	○	○	○	●
MAP	○	●	●	●	○	●	●	○	●	○	○	●
CMC + metacognition proposal	●	●	●	●	●	●	●	●	●	●	○	●
MERLIN2 + LLM	●	○	●	●	○	●	●	●	●	○	○	●
CASPER	●	●	●	●	○	●	○	○	●	○	●	○

Note. P = perception; At = attention; Act = action selection; M = memory; L = learning; R = reasoning; Pl = planning; MC = metacognition; C = communication; Mot = motivation/emotion; S = social cognition; NS = neuro-symbolic integration. Filled circles indicate explicit treatment in the cited architecture description; open circles indicate absent or insufficiently explicit treatment. The map synthesizes Laird (2012), Choi and Langley (2018), Ororbia and Kelly (2022), Safron (2020), Fabiano et al. (2024), Webb et al. (2025), Laird, Lebiere, and Rosenbloom (2017), Laird, Lebiere, Rosenbloom, and Stocco (2026), González-Santamarta, Rodríguez-Lera, Fernández-Llamas, et al. (2024), Sun (2024), and Vinanzi and Cangelosi (2024). Circle assignments reflect explicit treatment in the cited source descriptions; per-cell sources and coding rationale are provided as supplementary material (`table3_cell_evidence.csv`). The rows differ in analytic unit: implemented architectures or systems (e.g., ICARUS, CogNGen, SOFAI, MAP, MERLIN2 + LLM, CASPER), aggregated integration families that pool several published cases (e.g., Soar + Transformer, ACT-R + LLM cases, CLARION + LLM roles), and theoretical proposals not yet implemented at equal maturity (e.g., CMC + metacognition proposal, IWMT). A filled circle therefore denotes explicit treatment at the literature level and does not by itself distinguish a specified mechanism from an implemented or experimentally evaluated one; aggregated-family rows should not be read as equivalent single-system implementations.

4.6 Practical Application Landscape

Robotics remains the dominant application domain, but the nature of deployment has shifted from isolated cognitive demonstrations toward integrated service, industrial, and collaborative systems. MERLIN and MERLIN2 combine behavior generation, planning, monitoring, and natural-language interaction on ROS 2 (González-Santamarta, Rodríguez-Lera, Álvarez-Aparicio, et al. 2020). Other demonstrations connect task-general cognitive control and collaborative autonomy to embodied interaction (Cantucci and Falcone 2022; Frasca et al. 2020). CORTEX couples a cognitive robotics architecture with synchronized simulation and physical common-sense reasoning (Bustos et al. 2019; Trinidad Barnech et al. 2024). ArmarX makes memory a central bridge between symbolic planning and continuous sensorimotor representations (Peller-Konrad et al. 2023). CASPER and mind-inspired HRI architectures add explicit social perception, intention estimation, and adaptive interaction (Umbrico et al. 2022; Vinanzi and Cangelosi 2024). Recent socially assistive systems pair a cognitive architecture with explainable control to support trust (D'Arco et al. 2025).

Industrial and cyber-physical applications demonstrate concrete value. Soar-based semantic-memory reuse reduced automated laptop-disassembly time by more than 60% across models in one study (DiFilippo and Jouaneh 2019). Soar has also supported cooperative decisions in multi-unmanned-vehicle systems (Ding et al. 2024). Socially adaptive architectures have been developed for human-robot collaboration in industrial settings (Freire et al. 2024). LLM-ACTR and related work use cognitive architectures to improve grounding and explainability in manufacturing decisions (Wu, Souza, et al. 2024).

The application frontier is expanding beyond robotics. Proposed Soar-generative-AI combinations address cyberdefense and explainable recommendation (Prieto and Blakely 2024). Hybrid ACT-R memory has been used for personalization, and cognitive models support adaptive explanations and human-machine teaming (Lebiere, Cranford, et al. 2022). Cognitive simulation tools use LLMs to accelerate model authoring in ACT-R and Soar (Wu, Souza, et al. 2024). These cases suggest that a CA is most valuable when a system must maintain persistent state, expose reasons, adapt to a user, and recover from errors - requirements that are not unique to robots.

4.7 What Evidence Currently Supports

The evidence is strongest for the claim that architectural structure can improve a foundation model when a task requires persistent state, explicit decomposition, or constrained action. Bootstrapping a cognitive agent from LLM knowledge improved efficiency and generalization over an LLM-only agent in embodied kitchen tasks (Zhu and Simmons 2024). MAP improved LLM planning by separating and coordinating functions associated with conflict monitoring, state prediction, evaluation, decomposition, and task coordination (Webb et al. 2025). MERLIN2 experiments showed a complementary trade-off: the classical planner produced better plans, while the LLM-based solution improved natural interaction (González-Santamarta, Rodríguez-Lera, Fernández-Llamas, et al. 2024). Wilterson and Graziano (2021) demonstrated a controlled performance benefit from an explicit attention schema.

Other claims remain promising but under-tested. The CMC metacognition extension is a specification rather than a complete implementation (Laird, Lebiere, Rosenbloom, and Stocco 2026; Ukov and Tsochev 2025). Reviews of computational metacognitive architectures report highly heterogeneous terminology and limited quantitative evaluation; only 17% of architectures in one focused review were quantitatively evaluated for memory of metacognitive experience (Ukov and Tsochev 2025). Episodic-memory proposals are proliferating, but comparable tests of encoding, cueing, consolidation, and long-term use remain rare (Fountas et al. 2025). Social-cognition systems are generally evaluated in constrained scenarios, and uncertainty is seldom propagated in a principled way from learned representations to symbolic decisions.

5 An Architectural Governance Stack for Foundation-Model Agents

This section presents the paper's principal system-level synthesis, deriving the Architectural Governance Stack, its design requirements, and its interface contracts from the reviewed literature.

5.1 From Host Substrate to Governance Architecture

The literature supports a stronger synthesis than the observation that systems are becoming hybrid. Hybridization says that different computational paradigms coexist; it does not specify who controls whom, what information crosses interfaces, or how a system responds when components disagree. The recurring post-2018 requirement is governance. Foundation models generate high-value but fallible proposals. Persistent memories, symbolic constraints, uncertainty estimates, execution monitors, and human oversight determine whether those proposals become beliefs or actions (Kirk et al. 2024; Wray et al. 2025). The Architectural Governance Stack (AGS) in Figure 2 formalizes this relationship.

The processing loop begins with perception and multimodal grounding, continues through persistent state and memory, obtains one or more generative proposals, subjects them to explicit reasoning, planning, or verification, and executes an action or interaction. Three cross-cutting planes govern the loop. A metacognitive control plane manages uncertainty, provenance, trust, computational resources, safety, and human escalation. A lifelong-learning plane captures episodes, consolidates semantic knowledge, compiles skills, regulates intrinsic motivation, and limits forgetting. A social and normative layer represents other agents, roles, norms, consent, and collaborative expectations. The model is intentionally architecture-neutral: Soar (Laird 2012), ACT-R (Anderson 2007), CLARION (Sun 2024), CMC-derived systems (Laird, Lebiere, and Rosenbloom 2017), active-inference architectures (Friston et al. 2017), and newer LLM-agent systems (Sumers et al. 2024) may instantiate its functions differently.

The stack was constructed abductively from recurring cross-channel requirements in the corpus rather than imposed a priori. Table 6 makes this derivation explicit: for each layer it names the corpus evidence and contextual theory that motivated it, the recurring failure mode it is meant to address, and the resulting governance requirement. Section 5.2 compares the resulting stack with existing architectural frameworks.

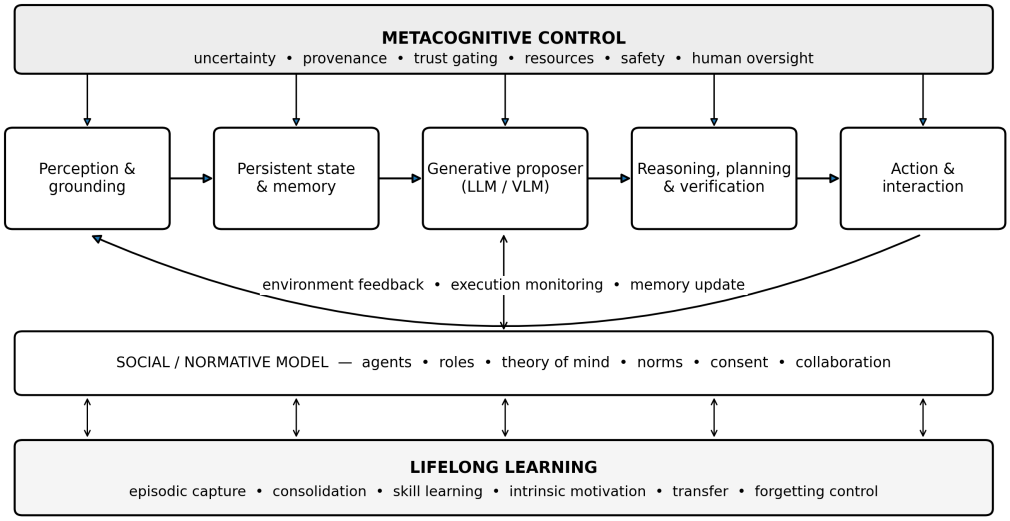


Fig. 2. Architectural Governance Stack. The model synthesizes recurrent functions in post-2018 cognitive architectures and positions foundation models as generative services inside metacognitive, memory, learning, and social control planes.

Table 6. Derivation of the Architectural Governance Stack: each layer linked to corpus evidence, contextual theory, the recurring failure it addresses, and the resulting governance requirement.

AGS layer	Corpus evidence	Contextual theory	Recurring failure addressed	Governance requirement
Perceptual grounding	Generative and multimodal perception in robot control loops (González-Santamarta, Rodríguez-Lera, Fernández-Llamas, et al. 2024)	Predictive processing / active inference (Friston et al. 2017)	Ungrounded labels without uncertainty	Update an explicit world model and expose uncertainty
Persistent state and memory	Episodic-memory-augmented and language-agent systems (Sumers et al. 2024)	Complementary learning systems (McClelland et al. 1995)	Statelessness between model calls; lost provenance	Complementary episodic/semantic memory with provenance
Generative proposal service	LLMs as proposers, planners, and knowledge sources (Kirk et al. 2024)	Dual-process cognition (Evans and Stanovich 2013)	Fluent but unverified or fabricated outputs	Treat foundation-model output as a proposal, not a belief

AGS layer	Corpus evidence	Contextual theory	Recurring failure addressed	Governance requirement
Reasoning, planning, verification	Verifier-gated reasoners and planner-plus-monitor systems (Fabiano et al. 2024; Webb et al. 2025)	Neuro-symbolic arbitration (Acharya and Song 2025)	Unchecked plans and silent execution errors	Explicit verification and arbitration before action
Metacognitive control plane	Computational metacognitive architectures (Laird, Lebiere, Rosenbloom, and Stocco 2026; T. Liu and van der Schaar 2025)	Nelson-Narens monitoring/control (Nelson and Narens 1990)	No calibration, escalation, or abstention	An uncertainty-, trust-, and escalation-aware controller
Lifelong-learning plane	Curiosity and consolidation mechanisms (Modirshanechi et al. 2023; Oudeyer and Kaplan 2007)	Intrinsic motivation; consolidation (McClelland et al. 1995)	Static knowledge; catastrophic forgetting	Episodic capture, consolidation, and controlled forgetting
Social and normative layer	Bounded theory-of-mind and HRI systems (Lebiere, Pirolli, et al. 2024; Z. Wang, Davis, et al. 2025)	Computational theory of mind (Jara-Ettinger 2019)	Unconstrained anthropomorphic inference	Bounded, revisable models of agents, roles, and norms
Typed interfaces (middle layer)	Interface-underspecification failures in integrated systems (Sarathy et al. 2025)	Provenance-aware neuro-symbolic exchange (Acharya and Song 2025)	Silent cross-module errors; non-reproducibility	Typed, provenance- and uncertainty-aware contracts with a veto path

5.2 Positioning against Existing Architectural Frameworks

Table 7 compares the Architectural Governance Stack with five frameworks that address overlapping problems: the CMC as the cross-lineage reference architecture (Laird, Lebiere, and Rosenbloom 2017); CoALA as the most direct precedent for language agents (Sumers et al. 2024); belief-desire-intention (BDI) architectures as the standard account of deliberative commitment (Rao and Georgeff 1995); the monitor-analyze-plan-execute loop over shared knowledge (MAPE-K) as the reference loop for autonomic self-management (Kephart and Chess 2003); and SOFAI as the closest precedent for metacognitive arbitration between fast and slow solvers (Fabiano et al. 2024). The comparison concerns architectural commitments as specified in the cited sources, not performance, and implementations of each framework may extend beyond its specification.

Three differences stand out. The first concerns the status of the foundation model. The CMC, classical BDI, and MAPE-K assign it no role, whereas CoALA places the language model at the centre of the decision cycle and typically relies on it to evaluate its own proposals. The stack treats the model as a fallible generative service whose outputs are proposals, not beliefs, plans, or actions. The second concerns the interface. None of the five frameworks requires a message crossing a module boundary to carry its semantic type, provenance, uncertainty,

and permitted downstream use; the stack makes these fields mandatory (Section 5.4) and gives verification an explicit veto. The third concerns the scope of arbitration. SOFAI already uses a metacognitive module to choose between fast and slow solvers on the basis of confidence and cost; the stack extends this control from solver selection to the admission of any generative output into memory, belief, or action, and places it alongside lifelong-learning and social-normative planes.

The stack is not intended to replace these frameworks. Its memory decomposition follows the CMC, its processing loop is compatible with the CoALA decision cycle, commitment to plans can be realized by BDI machinery, and supervisory adaptation by a MAPE-K loop. Each of these frameworks can therefore serve as a host for the stack; what the stack adds are the interface and arbitration commitments and the tests that can falsify them.

Table 7. Architectural Governance Stack compared with existing architectural frameworks, by architectural commitment.

Framework	**Role of learned or generative components**	**Memory and provenance**	**Arbitration and authorization**	**Learning, social layer, and evaluation**
CMC (Laird, Lebiere, and Rosenbloom 2017)	Not specified; formulated before foundation models. Later bridges place generative networks in peripheral modules (West et al. 2024)	Working, declarative, and procedural memory; provenance typing not specified	Procedural selection within a recurrent cognitive cycle; veto over module outputs not specified	All long-term knowledge learnable; social cognition not specified; assessed against human behavioral and neural data
CoALA (Sumers et al. 2024)	Language model as the central decision component	Working, episodic, semantic, and procedural modules; provenance not required	Propose–evaluate–select cycle, with evaluation typically by the language model; typed authorization not specified	Learning as actions that write to memory; social layer not specified; conceptual framework without an evaluation protocol
BDI (Rao and Georgeff 1995)	None in classical formulations	Belief base and plan library; provenance and calibrated uncertainty not specified	Commitment to intentions with reconsideration; adopted intentions are trusted	No learning plane in classical formulations; models of other agents not required; formal semantics
MAPE-K (Kephart and Chess 2003)	Not specified	Shared knowledge base; memory types not differentiated	Supervisory loop that monitors, analyzes, plans, and executes adaptations of a managed system	Adaptation policies; social-normative layer not specified; evaluated against self-management goals
SOFAI (Fabiano et al. 2024)	Fast solvers, which may be learned, alongside slow symbolic solvers	Models of the world and of the system's own solvers inform arbitration; provenance not required	Metacognitive selection between fast and slow solvers based on confidence and cost	Evaluated on planning tasks against individual solvers; social layer not central

Framework	Role of learned or generative components	Memory and provenance	Arbitration and authorization	Learning, social layer, and evaluation
AGS (this article)	Fallible generative service; outputs are proposals, not beliefs or actions	Complementary working, episodic, semantic, and procedural memory with source tags and provenance	Metacognitive plane with typed contracts, verification, and an explicit veto and escalation path before commitment	Lifelong-learning plane; bounded social-normative layer; minimum scorecard and falsifiable hypotheses (Table 10, Section 6.2)

5.3 Ten Design Requirements for Post-2018 Cognitive Architectures

The reviewed literature supports ten design requirements for cognitive architectures in the foundation-model era:

(1) *Generative perception with grounding.* Perception should update an explicit world model and expose uncertainty, rather than only emit labels (Friston et al. 2017).
(2) *An explicit model of attentional allocation.* The system should represent what information it is prioritizing and why, enabling control and diagnosis (Graziano and Webb 2015).
(3) *Complementary memory systems.* Fast episodic capture, slower semantic consolidation, working memory, procedural skill, deliberate cue construction, and controlled forgetting should be distinct but interoperable (McClelland et al. 1995).
(4) *Dual-process proposal and verification.* Fast neural or LLM proposals should be separable from slower symbolic or model-based evaluation (Evans and Stanovich 2013).
(5) *Metacognitive arbitration.* A controller should decide when to trust, verify, search, ask, abstain, or escalate, based on uncertainty and expected value of computation (Fabiano et al. 2024).
(6) *Language under architectural control.* Natural-language input and output should remain grounded in task state, memory provenance, and executable schemas (Kirk et al. 2024).
(7) *Intrinsic and extrinsic motivation.* Learning and exploration should balance task rewards, information gain, learning progress, and safety constraints (Oudeyer and Kaplan 2007).
(8) *Context-limited social cognition.* Theory of mind should be represented as testable, bounded models of relevant agents and roles rather than unconstrained anthropomorphic inference (Z. Wang, Davis, et al. 2025).
(9) *Typed neuro-symbolic interfaces.* Symbolic and subsymbolic components should exchange representations with specified semantics, calibration, provenance, and intervention behavior (Acharya and Song 2025).
(10) *Evaluation-ready implementation.* The architecture should log decisions, module calls, confidence, memory updates, failures, and resource costs so that its cognitive claims can be tested (Sarathy et al. 2025).

5.4 Interface Contracts: The Missing Middle Layer

Most integration failures occur not because either the neural or symbolic component is intrinsically inadequate, but because their interface is underspecified (Sarathy et al. 2025). An architecture should define at least six fields for every cross-module message: semantic type, source and timestamp, confidence or uncertainty representation, provenance, permissible downstream uses, and a repair or escalation route. For an LLM-generated plan, for example, the interface should distinguish a proposal from an approved plan; bind object names to current world-state entities; record the prompt, model, and retrieved evidence; attach a confidence or validation status; and prevent execution until constraints are satisfied. These contracts operationalize intervenability and make system behavior reproducible (Acharya and Song 2025).

Table 8. Practitioner decision guide for selecting an LLM-CA integration pattern.

Deployment condition	Recommended primary pattern	Why	Required safeguards	Avoid relying on
Open-ended dialogue with bounded actions	Language/multimodal interface plus symbolic execution	Maximizes linguistic flexibility while preserving action control	Grounding, clarification, schema validation, consent, and action confirmation	Unverified natural-language plans
Knowledge-intensive but auditable decisions	Knowledge/generative memory plus source-tagged semantic memory	Expands coverage while keeping evidence inspectable	Retrieval provenance, contradiction checks, recency controls, and abstention	Treating model recall as stored fact
Long-horizon planning in dynamic environments	Heuristic proposer plus model-based planner and execution monitor	Uses semantic decomposition without surrendering formal state tracking	Constraint checking, replanning, state reconciliation, and recovery tests	Single-pass plans or self-reported task success
Rapid domain or rule acquisition	Rule/model bootstrapping	Reduces manual knowledge engineering	Sandboxing, unit tests, version control, expert approval, and rollback	Direct insertion of generated rules into production
High-stakes adaptive reasoning	Bidirectional neuro-symbolic coupling under metacognitive control	Supports learning, explicit inference, and intervention	Typed interfaces, uncertainty translation, causal tests, audit logs, and human authority	Opaque end-to-end policies without inspection
Low-risk creative or exploratory task	Lightweight generative module with minimal architecture	Avoids unnecessary complexity	Basic monitoring, rate limits, and user-visible uncertainty	Over-engineering a full CA when persistent control is not needed

5.5 Illustrative Worked Example: A Modular Reasoning Architecture for Requirements Engineering

The framework can be instantiated concretely as a modular reasoning architecture for software requirements engineering (RE), a domain that exercises the governance problems raised above: long-lived project memory, multi-step deliberation, verification against a knowledge base, and normative constraints from standards and stakeholders. The architecture decomposes reasoning into four inference types (deduction, abduction, analogy, and induction) over a shared knowledge base, under a metacognitive orchestration layer, with a foundation model acting as a fallible generative service inside the control loop rather than as the reasoner itself (Proposition 2). This example illustrates, rather than proves, the paper's central claim that competence emerges from architectural governance of a generative service. The architecture-to-module assignments draw on Longo et al. (2025), Zi et al. (2025), Wu, Oltramari, et al. (2025), Jo and Kwon (2025), and Cheng et al. (2026).

Table 9. Architecture-to-module mapping for the proposed RE reasoning model.

Module	Inference type	Cognitive architectures	Ability channels	Governance role
Deduction	Formal / monotonic inference	SOAR + NARS	Reasoning	Verified inference
Abduction	Best-explanation hypotheses	CLARION + Sigma	Reasoning, learning	Hypothesis generation
Analogy	Structure mapping	ACT-R + LIDA	Reasoning, memory	Transfer / retrieval
Induction	Generalization	LIDA + CoALA	Reasoning, learning	Pattern abstraction
Context	Grounding via retrieval	CoALA + Graph-RAG	Memory, neuro-symbolic	Evidence grounding (KB)
Orchestration	Metacognitive arbitration	CLARION + BDI	Metacognition, action selection	Control & trust allocation

Architecture assignment is treated as a criteria-based, revisable design decision rather than a fixed commitment. Each candidate is scored against the evaluation dimensions of Section 6.1 (functional coverage of the target channel, evidence maturity, foundation-model integration readiness, RE fit, uncertainty handling, and openness), and the mapping illustrates a possible criteria-based selection procedure rather than a validated ranking; the full scoring rubric is left for empirical validation. Any later substitution, or the addition of a further architecture, is therefore a documented, auditable step in the spirit of an architecture decision record rather than an ad hoc change.

Controlled experiments on RE questions, previously reported by Sabaliauskas and Miliauskaite (2025) and cited here solely as external illustrative evidence, support this separation of concerns: stronger retrieval improved answer correctness for capable generation models but could not compensate for a very small model, which remained faithful to the retrieved context yet frequently incorrect. This indicates that the knowledge base is necessary but not sufficient, and motivates dedicated reasoning and orchestration modules, although those experiments do not by themselves establish their causal contribution. No new experimental results are reported in this review; the illustration is used to motivate, not to confirm, Propositions 2 and 4.

6 Research and Evaluation Agenda

This section translates the synthesis into a practitioner evaluation scorecard and a set of falsifiable hypotheses.

6.1 A Minimum Architecture Evaluation Scorecard

The field cannot compare architectures using task success alone, and recent work on LLM-robot integration similarly calls for evaluation that separates language robustness from explainability and verifiability (Sarathy et al. 2025). A system can complete a short task by exploiting an LLM while failing the architectural claims that motivated its design. Table 10 proposes a minimum scorecard that separates outcome performance from cognitive-system properties. The protocol is intended for comparative experiments and can be reported at system and module levels.

Table 10. Minimum evaluation scorecard for architecture-plus-foundation-model systems.

Dimension	Core question	Illustrative measures	Recommended comparison
Task effectiveness	Does the system complete the task correctly?	Success rate, quality, constraint violations, cumulative reward	Architecture versus LLM-only, symbolic-only, and ablated variants
Long-horizon reliability	Does performance persist over long sequences and repeated episodes?	Failure probability by horizon, recovery rate, loop frequency, false-success claims	Short versus long horizons; novel disturbances; repeated deployments
Memory and retention	Does memory encode, retrieve, consolidate, and forget appropriately?	Recall precision, encoding specificity, cue sensitivity, interference, retention, provenance accuracy	No memory; retrieval-only memory; episodic plus semantic consolidation
Transfer and lifelong learning	Does experience improve new tasks without destructive forgetting?	Forward transfer, backward transfer, sample efficiency, forgetting index	Frozen system; fine-tuning; episodic augmentation; metacognitive learning
Grounding and faithfulness	Are symbols, language, and actions tied to current evidence and state?	Entity-grounding accuracy, evidence attribution, executable-plan validity, explanation faithfulness	Ungrounded model; retrieval augmentation; symbolic verification
Uncertainty and metacognition	Does the system know when to verify, deliberate, ask, or abstain?	Calibration error, selective accuracy, monitoring resolution, control value, escalation quality	No metacognition; confidence-only; full monitoring-and-control layer
Safety and human oversight	Can failures be contained, reversed, and understood?	Unsafe-action rate, intervention success, rollback, policy compliance, audit completeness	No guard; static rules; metacognitive safety controller; human-in-the-loop
Efficiency and reproducibility	Are benefits worth the computational and engineering cost?	Latency, energy, token use, memory footprint, implementation effort, reproducibility package	Equal-resource baselines and component ablations

6.2 Priority Testable Hypotheses

The requirements-engineering illustration in Section 5.5 offers preliminary, indirect motivation for Propositions 2 and 4 (Sabaliauskas and Miliauskaite 2025), but no confirmatory evidence. The evidence gaps identified in this review support a focused experimental program:

H1 - Metacognitive control: adding a monitoring-and-control layer will improve selective accuracy, transfer, and recovery while reducing catastrophic forgetting and unsafe actions relative to an architecture-matched control without metacognition (Ukov and Tsochev 2025).

H2 - Episodic memory versus fine-tuning: source-tagged episodic memory with consolidation will outperform repeated fine-tuning for rapidly changing, long-horizon tasks when measured on retention, correction, and provenance, although fine-tuning may remain advantageous for stable skill acquisition.

H3 - Attention schema: an explicit model of the agent's own attention will improve allocation under distraction and partial observability beyond a strong attention mechanism that lacks self-description (Wilterson and Graziano 2021).

H4 - Dual-process arbitration: dual-process systems will dominate a single solver only when the controller is calibrated to task difficulty and resource cost; otherwise orchestration overhead will erase the benefit (Fabiano et al. 2024).

H5 - Bidirectional neuro-symbolic coupling: systems that transmit both learned proposals upward and symbolic constraints downward will show better compositional generalization and intervention success than one-directional hybrids (Wan et al. 2025).

H6 - Social cognition: bounded, context-specific theory-of-mind models will provide safer and more accurate HRI adaptation than unconstrained LLM inference about user beliefs and intentions (Z. Wang, Davis, et al. 2025).

H7 - Interface contracts: typed, provenance-aware interfaces will reduce silent cross-module errors and improve reproducibility more than adding model scale without interface redesign (Sarathy et al. 2025).

6.3 Reporting Standard for Cumulative Progress

Every new cognitive-architecture study should report: (a) the architectural diagram and module interfaces; (b) which ability claims are implemented rather than merely discussed; (c) training and external-model dependencies; (d) memory types and update rules; (e) uncertainty representation; (f) the controller that arbitrates components; (g) failure and recovery behavior; (h) latency and resource use; (i) all prompts, rules, ontologies, and configuration files needed to reproduce results; and (j) ablations that isolate the contribution of the architectural mechanism (González-Santamarta, Rodríguez-Lera, and Matellán-Olivera 2025; Sarathy et al. 2025). For LLM-integrated systems, reports should also distinguish model-generated proposals, accepted beliefs, approved plans, and executed actions (Kirk et al. 2024). This distinction is essential for both scientific validity and safety assurance.

7 Discussion

This section situates the findings within the broader literature, considers their implications, and highlights open problems.

7.1 Scientific Contribution: From Taxonomy to Governance

The principal scientific result is not that cognitive architectures have acquired more modules, but that the field is reorganizing around a layered theoretical account of governed hybrid agency. The historical architecture problem — how stable structures coordinate perception, memory, learning, reasoning, and action over time — remains, but familiar abilities are now re-instantiated on a substrate in which deep learned models are available by default. In each case the foundation model supplies proposals while the surrounding architecture supplies persistence, control, and verification.

This synthesis advances prior taxonomies in two related ways. The theoretical framework specifies what an implementation must represent, compute, coordinate, and empirically demonstrate before an ability label is warranted. The Architectural Governance Stack then makes governance the organizing relation among those abilities. In an AGS, an LLM is neither a mere peripheral tool nor the entire cognitive system. It is a generative service whose outputs can enter cognition only through explicit memory, reasoning, planning, social, and control interfaces. This framing avoids two symmetrical errors: treating classical architectures as obsolete because foundation models are broad, and treating an LLM as a complete architecture because prompts can imitate a decision loop. The key research question becomes whether the theoretical mechanism and its control substrate produce measurable gains in reliability, learning, and intervention.

7.2 Practical Contribution: Architecture Selection Becomes Risk-Based

For practitioners, the review suggests that a full cognitive architecture is not always warranted. Low-risk creative tasks may need only lightweight orchestration. Architecture becomes valuable as persistence, embodiment, accountability, or safety requirements increase. Systems that act in the world, remember users, update their own rules, or operate over long horizons require stronger separation between proposals, beliefs, plans, and actions (Kirk et al. 2024). The decision guide in Table 8 therefore links architectural investment to deployment risk rather than to an abstract aspiration toward human-like cognition.

The review also identifies practical design leverage. Existing ROS-based architectures can host LLM services without discarding established planners and monitors (González-Santamarta, Rodríguez-Lera, Fernández-Llamas, et al. 2024). Memory systems can be introduced before expensive model fine-tuning. Symbolic validators can constrain LLM plans while preserving natural interaction. Metacognitive policies can initially be implemented as explicit trust gates, escalation rules, and resource controllers, then replaced by learned arbitration once adequate evaluation data exist (T. Liu and van der Schaar 2025). This staged approach is more testable and auditable than end-to-end replacement.

7.3 Relationship to Previous Reviews

Kotseruba and Tsotsos (2020) established the breadth of the architecture landscape and the persistent difficulty of integration. The present review uses that baseline but changes the unit of analysis from architecture names to ability channels, interfaces, and governance roles. González-Santamarta, Rodríguez-Lera, and Matellán-Olivera (2025) provide the strongest focused evidence on autonomous robotics, hybrid adoption, ROS infrastructure, and evaluation practice; the current synthesis extends across language agents, memory, metacognition, social cognition, recommender systems, cyberdefense, and manufacturing. Neuro-symbolic reviews explain integration taxonomies and desirable properties (Acharya and Song 2025; Wan et al. 2025; W. Wang and Yang 2024); the AGS connects those properties to persistent cognitive functions and operational evaluation. Agentic-AI reviews describe rapidly changing frameworks (Abou Ali and Dornaika 2025); a recent book-length synthesis traces the broader computational evolution of the field (Kotseruba and Tsotsos 2025), while the present account supplies the ability-channel and governance perspective needed to interpret these developments.

7.4 Trust, Safety, and the Limits of Cognitive Analogy

Cognitive inspiration does not by itself make a system safe, human-like, or scientifically valid. Labels such as working memory, theory of mind, or metacognition should be reserved for mechanisms with explicit representations, update rules, and causal roles. An LLM statement about confidence is not necessarily calibrated monitoring; a vector store is not necessarily episodic memory; and a chain of text is not automatically deliberation. The scorecard therefore evaluates functional consequences and intervention behavior rather than terminology.

At the same time, cognitive architectures offer safety-relevant affordances that end-to-end models lack: inspectable state, modular isolation, explicit goals and constraints, recovery policies, and the possibility of human intervention at defined control points (Acharya and Song 2025). These affordances become meaningful only when implemented and tested. A symbolic verifier that is bypassed under latency pressure, a memory without provenance, or a metacognitive controller trained on the same biases as its base model provides an appearance of governance rather than governance itself. High-stakes applications require adversarial testing of the whole control loop (Sarathy et al. 2025).

8 Limitations

This review has six principal limitations. First, the database query required foundation- or generative-model terminology. The resulting corpus is therefore suitable for mapping foundation-model-related cognitive architectures but cannot establish the prevalence of mechanisms across all post-2018 cognitive-architecture research.

Second, the corpus was produced through a reproducible rule-based pre-filter followed by an LLM-assisted title-and-abstract assessment, and not every candidate record was independently human-rescreened. Although a second model pass showed high repeatability (Cohen's $\kappa = 0.90$ on a random 50-record sample; Section 3.6 and Table 3), the purposive human audit identified five potentially eligible records that the model excluded and two records it retained that the author would have excluded (Section 3.6). The evidence map may therefore contain both false exclusions and false inclusions, and its frequencies should be interpreted descriptively rather than as unbiased estimates.

Third, English-language Scopus and Web of Science indexing excludes preprints, books, technical reports, non-English research, and systems documented mainly in software repositories; the ACM Digital Library, IEEE Xplore, and PsycINFO were also outside scope. The few non-core citations retained for framework context are indexed conference or journal versions.

Fourth, ability-channel and integration-pattern assignments require interpretive judgement. Even with an explicit codebook and reliability checks, complex systems may instantiate mechanisms only partially, distribute them across modules, or use cognitive terminology without causal implementation; a module marked explicit was one the cited description treated as an architectural component, which does not establish equivalent computational depth or performance across systems.

Fifth, the evidence base is heterogeneous and often evaluates task performance rather than architectural mechanisms. Implementation maturity does not itself establish validity, and many studies lack matched ablations, calibration, long-horizon evaluation, disturbance recovery, or reproducible artifacts; consequently no pooled effect size or formal quality score was possible.

Sixth, the Architectural Governance Stack, propositions, interface contracts, and scorecard are integrative proposals derived from recurring requirements in the literature. They are not empirical proof of a universal architecture and require prospective testing across domains, risks, and deployment conditions. Some anchors, such as the CMC (Laird, Lebiere, and Rosenbloom 2017) and complementary learning systems (McClelland et al. 1995), have extensive histories, whereas others, such as Attention Schema Theory in artificial agents (Wilterson and Graziano 2021) and intrinsic metacognitive learning (T. Liu and van der Schaar 2025), have narrower implementation evidence. Relatedly, the architecture assignment in the worked requirements-engineering example (Section 5.5) is grounded in the reviewed literature and preliminary experiments, is explicitly revisable, and its module architectures may be substituted or supplemented as evaluation-scorecard evidence accumulates.

These limitations define the appropriate claim: the article offers a focused, auditable evidence map and a falsifiable organizing framework for governed foundation-model agency, not an exhaustive taxonomy or a definitive account of cognition.

9 Conclusion

Foundation models expand the competence available to artificial agents, but they do not remove the architecture problem. Persistent agency still requires explicit commitments about state, memory, learning, action authorization, uncertainty, verification, recovery, and social or normative constraints. The focused evidence map reviewed here identifies five recurring ways in which foundation models are placed inside architecture-level systems and shows that the most consequential design choices lie at the interfaces: what information is accepted as a belief, how provenance and uncertainty are preserved, when slower verification overrides a generative proposal, how execution failure changes subsequent control, and which experiences are retained or forgotten. The theoretical

reconstruction shows continuity across this transition: the CMC preserves a common functional skeleton; predictive and active-inference traditions connect perception, action, attention, and learning; Attention Schema Theory adds self-modelled allocation; complementary systems explain episodic and semantic memory; curiosity drives open-ended learning; dual-process organization separates proposal from verification; and computational theory of mind introduces bounded models of other agents.

The governed-hybrid-agency framework and Architectural Governance Stack synthesize these recurring requirements into falsifiable comparison targets. Their scientific value will depend on intervention-sensitive studies that compare matched systems with and without persistent memory, typed interfaces, calibrated arbitration, verification, and recovery. Their practical value will depend on risk-based evaluation across grounding, calibration, retention, transfer, safety, latency, cost, and reproducibility. The central conclusion is therefore bounded: foundation models are most promising as powerful but fallible services within persistent, inspectable control systems. Whether a particular architecture realizes governed agency must be demonstrated through causal evidence, not inferred from cognitive terminology or component count.

Acknowledgments

We thank Dr. Jolanta Miliauskaitė (Institute of Data Science and Digital Technologies, Faculty of Mathematics and Informatics, Vilnius University) for her contribution to the revision of this article, in particular the comparative analysis of governed hybrid agency against existing architectural frameworks and the accompanying comparison table (Section 5.2 and Table 7), and for comments that helped sharpen the statement of the framework's novelty.

Use of Generative Artificial Intelligence

The review workflow combined four distinct steps: (1) a deterministic, rule-based pre-filter; (2) an LLM-assisted title-and-abstract eligibility assessment; (3) human full-text consultation of the included works; and (4) author synthesis. Generative AI (a large language model, Anthropic Claude) was used under author supervision for step (2), for structured metadata extraction, and for language editing. The LLM-assisted assessment determined the provisional title-and-abstract eligibility set that defines the 119-work corpus; these decisions were not comprehensively and independently human-validated, so the final corpus is treated as a machine-assisted focused evidence map. The authors defined all eligibility and coding criteria, consulted the full texts of every included work, conducted the synthesis, and take full responsibility for the content. The model, its settings, the exact prompt, and the per-record decisions are provided in the supplementary material.

Author Contributions

D.S. designed and executed the search, screening, extraction, analysis, and evidence mapping. D.S. and A.D. conceptualized the framework, interpreted the findings, and wrote and revised the manuscript.

Data Availability

The review evidence table, search strings, screening decisions, the combined 30-record calibration table (rule-based pre-filter, model, and author decisions), the 50-record reproducibility validation, the maturity and integration-pattern coding of the corpus, the language-model prompt and settings, the per-record maturity and abstract-level mechanism coding, and the coverage-matrix cell evidence are provided as supplementary material with this submission. The full package is available at https://github.com/DariusSabaliauskasVU/governed-hybrid-agency. To provide a stable reference, the package is distributed as a tagged release (v1.0.1) of the repository above, and that release is archived on Zenodo under DOI 10.5281/zenodo.23282942; all analyses in this article correspond to that release.

References

M. Abou Ali and F. Dornaika. 2025. "Agentic AI: A comprehensive survey of architectures, applications, and future directions." *Artificial Intelligence Review*. doi:10.1007/s10462-025-11422-4.

K. Acharya and H. Song. 2025. "A comprehensive review of neuro-symbolic AI for robustness, uncertainty quantification, and intervenability." *Arabian Journal for Science and Engineering*. doi:10.1007/s13369-025-10887-3.

P. Allgeuer, H. Ali, and S. Wermter. 2024. "When robots get chatty: Grounding multimodal human-robot conversation and collaboration." In: *Lecture Notes in Computer Science. Springer*. doi:10.1007/978-3-031-72341-4_21.

J. R. Anderson. 2007. *How can the human mind occur in the physical universe?* Oxford University Press.

L. Bärmann, F. Peller-Konrad, S. Constantin, T. Asfour, and A. Waibel. 2021. "Deep episodic memory for verbalization of robot experience." *IEEE Robotics and Automation Letters*, 6, 5809–5816. doi:10.1109/LRA.2021.3085166.

A. Brasoveanu and J. Dotlačil. 2021. "Reinforcement learning for production-based cognitive models." *Topics in Cognitive Science*. doi:10.1111/tops.12546.

P. Bustos, L. J. Manso, A. Bandera, J. Rubio, I. García-Varea, and J. Martínez-Gómez. 2019. "The CORTEX cognitive robotics architecture: Use cases." *Cognitive Systems Research*, 55, 107–123.

F. Cantucci and R. Falcone. 2022. "Collaborative autonomy: Human-robot interaction to the test of intelligent help." *Electronics*, 11, 19, 3065. doi:10.3390/electronics11193065.

Y. Cheng, Y. Zhao, J. Zhu, Y. Liu, X. Sun, and X. Li. 2026. "Human cognition inspired RAG with knowledge graph for complex problem solving." *Proceedings of the AAAI Conference on Artificial Intelligence*, 40, 36. doi:10.1609/aaai.v40i36.40291.

D. Choi. 2024. "On using generative models in a cognitive architecture for embodied agents." In: *Proceedings of the AAAI Symposium Series, 2(1)*. doi:10.1609/aaaiss.v2i1.27684.

D. Choi and P. Langley. 2018. "Evolution of the ICARUS cognitive architecture." *Cognitive Systems Research*, 48, 25–38. doi:10.1016/j.cogsys.2017.05.005.

A. Ciria, G. Schillaci, G. Pezzulo, V. V. Hafner, and B. Lara. 2021. "Predictive processing in cognitive robotics: A review." *Neural Computation*, 33, 5, 1402–1432. doi:10.1162/neco_a_01383.

B. Conway-Smith and R. L. West. 2022. "System-1 and System-2 realized within the Common Model of Cognition." In: *Proc. AAAI 2022 Fall Symposium: Thinking Fast and Slow and Other Cognitive Theories in AI (CEUR-WS Vol. 3332)*.

L. D'Arco, L. Raggioli, G. Randazzo, G. De Gasperis, A. Chella, S. Costantini, and S. Rossi. 2025. "Towards trustworthy and explainable socially assistive robots: A cognitive architecture for dietary guidance." In: *2025 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*. doi:10.1109/SIMPAR62925.2025.10979036.

C. L. Dancy and D. Workman. 2024. "On integrating generative models into cognitive architectures for improved computational sociocultural representations." In: *Proceedings of the AAAI Symposium Series, 2(1)*. doi:10.1609/aaaiss.v2i1.27685.

N. DiFilippo and M. Jouaneh. 2019. "Using the Soar cognitive architecture to remove screws from different laptop models." *IEEE Transactions on Automation Science and Engineering*, 16, 4, 1680–1690.

N. DiGilio and P. Agrawal. 2026. "Neuro-symbolic LIDA's semantic vision system." In: *Lecture Notes in Computer Science*. doi:10.1007/978-3-032-00686-8_14.

L. Ding, Y. Tang, T. Wang, T. Xie, P. Huang, and B. Yang. 2024. "A cooperative decision-making approach based on a Soar cognitive architecture for multi-unmanned vehicles." *Drones*, 8, 4, 155. doi:10.3390/drones8040155.

E. Erdoğan, F. Dignum, R. Verbrugge, and P. Yolum. 2025. "TOMA: Computational theory of mind with abstractions for hybrid intelligence." *Journal of Artificial Intelligence Research*, 82, 285–311. doi:10.1613/jair.1.16402.

J. S. B. T. Evans and K. E. Stanovich. 2013. "Dual-process theories of higher cognition: Advancing the debate." *Perspectives on Psychological Science*, 8, 3, 223–241. doi:10.1177/1745691612460685.

F. Fabiano, V. Pallagani, M. B. Ganapini, A. Loreggia, K. Murugesan, F. Rossi, and B. Srivastava. 2024. "Plan-SOFAI: A neuro-symbolic planning architecture." In: *AAAI 2024 Workshop on Neuro-Symbolic Learning and Reasoning*.

Z. Fountas, M. A. Benfeghoul, A. Oomerjee, F. Christopoulou, G. Lampouras, H. Bou-Ammar, and J. Wang. 2025. "Human-inspired episodic memory for infinite context LLMs." In: *13th International Conference on Learning Representations (ICLR 2025)*, 22411–22449.

T. M. Frasca, Z. Han, H. Yanco, and M. Scheutz. 2020. "Going cognitive: A demonstration of the utility of task-general cognitive architectures for adaptive robotic task performance." In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*.

I. T. Freire, O. Guerrero-Rosado, A. F. Amil, and P. Verschure. 2024. "Socially adaptive cognitive architecture for human-robot collaboration in industrial settings." *Frontiers in Robotics and AI*, 11, 1248646. doi:10.3389/frobt.2024.1248646.

K. J. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo. 2017. "Active inference: A process theory." *Neural Computation*, 29, 1, 1–49. doi:10.1162/NECO_a_00912.

M. Á. González-Santamarta, F. J. Rodríguez-Lera, C. Álvarez-Aparicio, Á. Guerrero-Higueras, and C. Fernández-Llamas. 2020. "MERLIN: A cognitive architecture for service robots." *Applied Sciences*, 10, 17, 5989. doi:10.3390/app10175989.

M. Á. González-Santamarta, F. J. Rodríguez-Lera, C. Fernández-Llamas, and V. Matellán-Olivera. 2024. "A hybrid cognitive architecture to generate, control, plan, and monitor behaviors for interactive autonomous robots." *International Journal of Social Robotics*. doi:10.1007/s12369-024-01192-4.

M. Á. González-Santamarta, F. J. Rodríguez-Lera, and V. Matellán-Olivera. 2025. "Cognitive architectures in autonomous robotics: A systematic review of behavior generation approaches and evaluation strategies." *IEEE Access*, 13, 191619–191644. doi:10.1109/ACCESS.2025.3630344.

M. S. A. Graziano and T. W. Webb. 2015. "The attention schema theory: A mechanistic account of subjective awareness." *Frontiers in Psychology*, 6, 500. doi:10.3389/fpsyg.2015.00500.

H. Hake, C. Sibert, and A. Stocco. 2022. "Inferring a cognitive architecture from multitask neuroimaging data: A data-driven test of the Common Model of Cognition using Granger causality." *Topics in Cognitive Science*. doi:10.1111/tops.12623.

W. Huang, A. Chella, and A. Cangelosi. 2024. "A cognitive robotics implementation of global workspace theory for episodic memory interaction with consciousness." *IEEE Transactions on Cognitive and Developmental Systems*, 16, 936–949. doi:10.1109/TCDS.2023.3266103.

J. Jara-Ettinger. 2019. "Theory of mind as inverse reinforcement learning." *Current Opinion in Behavioral Sciences*, 29, 105–110. doi:10.1016/j.cobeha.2019.04.010.

R. Jian, Y. Shang, Z. Zhu, K. Liu, C. Bao, and X. Liu. 2025. "CLMC: Code-level metacognition for LLM agents." In: *IEEE International Conference on Data Mining Workshops (ICDMW)*. doi:10.1109/ICDMW69685.2025.00295.

H.-R. Jo and Y.-W. Kwon. 2025. "Cognitive consistency verification: A neuro-symbolic approach for mitigating VLM hallucination." In: *Proceedings of the IEEE International Conference on Big Data (Big Data)*. doi:10.1109/BigData66926.2025.11402484.

H. Joshi and V. Ustun. 2024. "Augmenting cognitive architectures with large language models." In: *Proceedings of the AAAI Symposium Series, 2(1)*. doi:10.1609/aaaiss.v2i1.27689.

J. O. Kephart and D. M. Chess. 2003. "The vision of autonomic computing." *Computer*, 36, 1, 41–50. doi:10.1109/MC.2003.1160055.

J. R. Kirk, R. E. Wray, and J. E. Laird. 2024. "Exploiting language models as a source of knowledge for cognitive agents." In: *Proc. AAAI Symposium Series, 2(1), 286-294.* doi:10.1609/aaaiss.v2i1.27690.

I. Kotseruba and J. K. Tsotsos. 2020. "40 years of cognitive architectures: Core cognitive abilities and practical applications." *Artificial Intelligence Review*, 53, 1, 17–94. doi:10.1007/s10462-018-9646-y.

I. Kotseruba and J. K. Tsotsos. 2025. *The computational evolution of cognitive architectures*. Oxford University Press. doi:10.1093/oso/9780192844835.001.0001.

J. E. Laird. 2012. *The Soar cognitive architecture*. MIT Press.

J. E. Laird, C. Lebiere, and P. S. Rosenbloom. 2017. "A Standard Model of the Mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics." *AI Magazine*, 38, 4, 13–26. doi:10.1609/aimag.v38i4.2744.

J. E. Laird, C. Lebiere, P. S. Rosenbloom, and A. Stocco. 2026. "Unified, Comprehensive Metacognition within the Common Model of Cognition." In: *Artificial General Intelligence: 19th International Conference, AGI 2026* (Lecture Notes in Artificial Intelligence). Ed. by M. Iklé et al. Vol. 16855. Springer Nature Switzerland, Cham, 1–15. doi:10.1007/978-3-032-33195-3_1.

J. E. Laird, R. E. Wray, S. Jones, J. R. Kirk, and P. Lindes. 2024. "Proposal for cognitive architecture and Transformer integration: Online learning from agent experience." In: *Proceedings of the AAAI Symposium Series, 2(1)*. doi:10.1609/aaaiss.v2i1.27692.

C. Lebiere, E. A. Cranford, M. Martin, D. Morrison, and A. Stocco. 2022. "Cognitive architectures and their applications." In: *2022 IEEE International Conference on Communications in Computing (pp. 94-99)*. doi:10.1109/CIC56439.2022.00018.

C. Lebiere, P. Pirolli, M. Johnson, M. Martin, and D. Morrison. 2024. "Cognitive models for machine theory of mind." *Topics in Cognitive Science*. doi:10.1111/tops.12773.

A. Lieto, M. Bhatt, A. Oltramari, and D. Vernon. 2018. "The role of cognitive architectures in general artificial intelligence." *Cognitive Systems Research*, 48, 1–3. doi:10.1016/j.cogsys.2017.08.003.

T. Liu and M. van der Schaar. 2025. "Position: Truly self-improving agents require intrinsic metacognitive learning." *Proceedings of Machine Learning Research*, 267, 81714–81727.

C. F. Longo, M. Mongiovì, L. Bulla, and A. Lieto. 2025. "Eliciting metaknowledge in large language models." *Cognitive Systems Research*, 91, 101352. doi:10.1016/j.cogsys.2025.101352.

J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly. 1995. "Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory." *Psychological Review*, 102, 3, 419–457. doi:10.1037/0033-295X.102.3.419.

A. Modirshanechi, K. Kondrakiewicz, W. Gerstner, and S. Haesler. 2023. "Curiosity-driven exploration: Foundations in neuroscience and computational modeling." *Trends in Neurosciences*, 46, 12, 1054–1069. doi:10.1016/j.tins.2023.10.002.

S. Mohan, M. Klenk, M. Shreve, K. Evans, A. Ang, and J. Maxwell. 2020. "Characterizing an analogical concept memory for architectures implementing the Common Model of Cognition." *Advances in Cognitive Systems*, 8.

V. Nadimuthu, A. Das, V. Soni, and A. P. Sheth. 2026. "Symbolic-in-the-loop: Leveraging knowledge graph engineering to ensure high-assurance generative artificial intelligence outcomes." *IEEE Internet Computing*. doi:10.1109/MIC.2026.3680545.

K. Nagashima, J. Morita, and Y. Takeuchi. 2024. "Intrinsic motivation in cognitive architecture: Intellectual curiosity originated from pattern discovery." *Frontiers in Artificial Intelligence*, 7, 1397860. doi:10.3389/frai.2024.1397860.

T. Navarro. 2025. "Enhancing cognitive functions in large language models towards AGI: A state-of-the-art survey and exploratory review." In: *2025 10th International Conference on Machine Learning Technologies (ICMLT)*. doi:10.1109/ICMLT65785.2025.11193354.

T. O. Nelson and L. Narens. 1990. "Metamemory: A theoretical framework and new findings." *G. H. Bower (Ed.), The psychology of learning and motivation (Vol. 26, pp. 125-173). Academic Press*. doi:10.1016/S0079-7421(08)60053-5.

A. Newell. 1990. *Unified theories of cognition*. Harvard University Press.

A. Ororbia and M. A. Kelly. 2024. "A neuro-mimetic realization of the Common Model of Cognition via Hebbian learning and free energy minimization." In: *Proc. AAAI Symposium Series, 2(1), 369-378*. doi:10.1609/aaaiss.v2i1.27702.

A. Ororbia and M. A. Kelly. 2022. "CogNGen: Constructing the kernel of a hyperdimensional predictive processing cognitive architecture." In: *Proc. 44th Annual Conf. Cognitive Science Society (pp. 1322-1329)*.

P.-Y. Oudeyer and F. Kaplan. 2007. "What is intrinsic motivation? A typology of computational approaches." *Frontiers in Neurorobotics*, 1, 6. doi:10.3389/neuro.12.006.2007.

M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, and D. Moher. 2021. "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews." *BMJ*, 372, n71. doi:10.1136/bmj.n71.

A. E. Pecora, F. Strada, and A. Bottino. 2026. "A survey of memory models for virtual agents and humans: From psychological foundations to computational architectures." In: *Lecture Notes in Computer Science*. doi:10.1007/978-3-031-97778-7_8.

F. Peller-Konrad, R. Kartmann, C. R. G. Dreher, A. Meixner, F. Reister, M. Grotz, and T. Asfour. 2023. "A memory system of a robot cognitive architecture and its implementation in ArmarX." *Robotics and Autonomous Systems*, 164, 104415. doi:10.1016/j.robot.2023.104415.

G. Pezzulo, T. Parr, and K. Friston. 2022. "The evolution of brain architectures for predictive coding and active inference." *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377, 1844, 20200531. doi:10.1098/rstb.2020.0531.

D. Premack and G. Woodruff. 1978. "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, 1, 4, 515–526. doi:10.1017/S0140525X00076512.

T. J. Prescott and P. F. Dominey. 2024. "Synthesizing the temporal self: Robotic models of episodic and autobiographical memory." *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379, 20230415. doi:10.1098/rstb.2023.0415.

I. Prieto and B. Blakely. 2024. "Proposed uses of generative AI in a cybersecurity-focused Soar agent." In: *Proceedings of the AAAI Symposium Series, 2(1)*. doi:10.1609/aaaiss.v2i1.27704.

R. Proietti, T. Parr, A. Tessari, K. Friston, and G. Pezzulo. 2025. "Active inference and cognitive control: Balancing deliberation and habits through precision optimization." *Physics of Life Reviews*, 54, 27–51. doi:10.1016/j.plrev.2025.05.008.

A. S. Rao and M. P. Georgeff. 1995. "BDI agents: From theory to practice." In: *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*. AAAI Press, 312–319.

A. Romero, F. Bellas, and R. J. Duro. 2023. "A perspective on lifelong open-ended learning autonomy for robotics through cognitive architectures." *Sensors*, 23, 3, 1611. doi:10.3390/s23031611.

O. J. Romero, J. Zimmerman, A. Steinfeld, and A. Tomasic. 2023. "Synergistic integration of large language models and cognitive architectures for robust AI: An exploratory analysis." In: *Proceedings of the AAAI Symposium Series*.

D. Sabaliauskas and J. Miliauskaite. 2025. "Enhancing requirements engineering QA with retrieval-augmented generation and knowledge graphs." In: *Vilnius University Proceedings (DAMSS-2025)*. doi:10.15388/DAMSS.16.2025.

A. Safron. 2020. "An Integrated World Modeling Theory (IWMT) of consciousness." *Frontiers in Artificial Intelligence*, 3, 30. doi:10.3389/frai.2020.00030.

V. Sarathy, M. Fawn, M. McWilliams, M. Scheutz, and B. Oosterveld. 2025. "On evaluating LLM integration into robotic architectures." *ACM Transactions on Intelligent Systems and Technology*. doi:10.1145/3754340.

M. Scheibl, B. Richter, A. Müller, M. Beetz, and B. Wrede. 2025. "Towards a cognitive architecture to enable natural language interaction in co-constructive task learning." In: *2025 IEEE International Conference on Robot and Human Interactive Communication*. doi:10.1109/RO-MAN63969.2025.11217812.

T. Sievers and N. Russwinkel. 2025. "Retrieving memory content from a cognitive architecture by impressions from language models for use in a social robot." *Applied Sciences*, 15, 10, 5778. doi:10.3390/app15105778.

T. R. Sumers, S. Yao, K. Narasimhan, and T. L. Griffiths. 2024. "Cognitive architectures for language agents." *Transactions on Machine Learning Research*. https://openreview.net/forum?id=1i6ZCvflQJ.

R. Sun. 2024. "Dual-process theories, cognitive architectures, and hybrid neural-symbolic models." *Neurosymbolic Artificial Intelligence*, 1, 115–125. doi:10.3233/NAI-240720.

A. Ten, P.-Y. Oudeyer, and C. Moulin-Frier. 2022. *Curiosity-driven exploration: Diversity of mechanisms and functions*. The drive for knowledge (pp. 53-76). Cambridge University Press.

G. Trinidad Barnech, G. Tejera, J. Valle-Lisboa, P. Núñez Trujillo, P. Bachiller, and P. Bustos. 2024. "Enhancing robotic perception through synchronized simulation and physical common-sense reasoning." *Sensors*, 24, 7, 2249. doi:10.3390/s24072249.

T. Ukov and G. Tsochev. 2025. "Reviewing a model of metacognition for application in cognitive architecture design." *Systems*, 13, 3, 177. doi:10.3390/systems13030177.

A. Umbrico, R. De Benedictis, F. Fracasso, A. Cesta, A. Orlandini, and G. Cortellessa. 2022. "A mind-inspired architecture for adaptive HRI." *International Journal of Social Robotics*, 14, 1343–1360. doi:10.1007/s12369-022-00897-8.

S. Vinanzi and A. Cangelosi. 2024. "CASPER: Cognitive architecture for social perception and engagement in robots." *International Journal of Social Robotics*. doi:10.1007/s12369-024-01116-2.

Z. Wan, H. Yang, J. Qian, R. Raj, J. Park, C. Wang, A. Raychowdhury, and T. Krishna. 2025. "Compositional AI beyond LLMs: System implications of neuro-symbolic-probabilistic architectures." In: *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. doi:10.1145/3760250.3762235.

W. Wang and Y. Yang. 2024. "Towards data- and knowledge-driven AI: A survey on neuro-symbolic computing." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi:10.1109/TPAMI.2024.3483273.

Z. Wang, I. Davis, and J. Jara-Ettinger. 2025. "Modeling other minds: A computational account of social cognition and its development." *Annual Review of Developmental Psychology*. doi:10.1146/annurev-devpsych-111323-112016.

Z. Wang, M. Gong, J. Xu, Q. Sun, S. Mao, X. Jin, Q.-L. Han, and Y. Tang. 2026. "Multi-agent systems: From classical paradigms to large foundation model-enabled futures." *IEEE/CAA Journal of Automatica Sinica*, 13, 5, 1007–1023. doi:10.1109/JAS.2026.126113.

T. W. Webb, S. Mondal, and I. Momennejad. 2025. "A brain-inspired agentic architecture to improve planning with LLMs." *Nature Communications*, 16, 8633. doi:10.1038/s41467-025-63804-5.

R. L. West, S. Eckler, B. Conway-Smith, N. Turcas, E. Tomkins-Flanagan, and M. A. Kelly. 2024. "Bridging generative networks with the Common Model of Cognition." In: *Proceedings of the AAAI Symposium Series, 2(1)*. doi:10.1609/aaaiss.v2i1.27709.

R. Whittemore and K. Knafl. 2005. "The integrative review: Updated methodology." *Journal of Advanced Nursing*, 52, 5, 546–553. doi:10.1111/j.1365-2648.2005.03621.x.

C. J. Whyte and R. Smith. 2021. "The predictive global neuronal workspace: A formal active inference model of visual consciousness." *Progress in Neurobiology*, 199, 101918. doi:10.1016/j.pneurobio.2020.101918.

A. I. Wilterson and M. S. A. Graziano. 2021. "The attention schema theory in a neural network agent: Controlling visuospatial attention using a descriptive model of attention." In: *Proceedings of the National Academy of Sciences, 118(33), e2102421118*. doi:10.1073/pnas.2102421118.

R. E. Wray, J. R. Kirk, and J. E. Laird. 2025. "Applying cognitive design patterns to general LLM agents." In: *Artificial General Intelligence (AGI 2025), LNAI 16058. Springer*. doi:10.1007/978-3-032-00800-8_28.

S. Wu, R. F. Souza, F. E. Ritter, and W. T. Lima Jr. 2024. "Comparing LLMs for prompt-enhanced ACT-R and Soar model development: A case study in cognitive simulation." In: *Proceedings of the AAAI Symposium Series, 2(1)*. doi:10.1609/aaaiss.v2i1.27710.

S. Wu, A. Oltramari, and F. E. Ritter. 2025. "VSM-ACTR 2: A human-like decision-making model with metacognition for manufacturing solutions." *Computational and Mathematical Organization Theory*. doi:10.1007/s10588-025-09405-5.

Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, and T. Gui. 2025. "The rise and potential of large language model based agents: A survey." *Science China Information Sciences*, 68, 2, 121101. doi:10.1007/s11432-024-4222-0.

Z. Zhang, Q. Dai, X. Bo, C. Ma, R. Li, X. Chen, J. Zhu, Z. Dong, and J.-R. Wen. 2025. "A survey on the memory mechanism of large language model-based agents." *ACM Transactions on Information Systems*, 43, 6, 1–47. doi:10.1145/3748302.

Y. Zhou, Y. Liu, X. Li, Y. Zhao, X. Wang, J. Tian, Z. Li, and X. Xu. 2025. "Metagent-P: A neuro-symbolic planning agent with metacognition for open worlds." In: *Findings of the Association for Computational Linguistics: ACL 2025*. doi:10.18653/v1/2025.findings-acl.1169.

Y. Zhou, Z. Liu, J. Jin, J.-Y. Nie, and Z. Dou. 2024. "Metacognitive retrieval-augmented large language models." In: *Proceedings of the ACM Web Conference 2024 (WWW '24)*. doi:10.1145/3589334.3645481.

F. Zhu and R. Simmons. 2024. "Bootstrapping cognitive agents with a large language model." In: *Proceedings of the AAAI Conference on Artificial Intelligence, 38(1), 620-628*. doi:10.1609/aaai.v38i1.27822.

Y. Zi, K. Roy, and A. Sheth. 2025. "Cognitive neurosymbolic artificial intelligence for complex decision-making: Integrating foundation models, cognitive architectures, and knowledge." *IEEE Intelligent Systems*. doi:10.1109/MIS.2025.3561487.

A Search Strings

Scopus (Advanced search, executed 10 July 2026): TITLE-ABS-KEY (("cognitive architecture*" OR "cognitive system*" OR "unified theor* of cognition" OR "ACT-R" OR "Soar" OR "Sigma" OR "LIDA" OR "CLARION" OR "cognitive agent*") AND ("large language model*" OR "foundation model*" OR "vision-language model*" OR "generative model*" OR "LLM" OR "neuro-symbolic" OR "neurosymbolic" OR "transformer*") AND (agent* OR reasoning OR planning OR metacognit* OR memory OR grounding OR governance OR verification OR autonom*)) AND PUBYEAR > 2017, limited to English-language articles, reviews, and conference papers.

Web of Science (Core Collection, Topic = TS; executed 10 July 2026): TS=(("cognitive architecture*" OR "cognitive system*" OR "unified theor* of cognition" OR "ACT-R" OR "Soar" OR "LIDA" OR "CLARION" OR "cognitive agent*") AND ("large language model*" OR "foundation model*" OR "vision-language model*" OR "generative model*" OR "neuro-symbolic" OR "neurosymbolic" OR "LLM") AND (agent* OR reasoning OR planning OR metacognit* OR memory OR grounding OR governance OR verification OR autonom*)), refined to 2018-2026, document types Article/Review/Proceedings Paper, English. Database-specific query adaptations (for example, the Scopus-only terms Sigma and transformer*) accommodated differences in indexing and query behaviour; the conceptual eligibility blocks (A, B, and C) were identical across the two databases.

B Supplementary Tables

Tables S1–S4 provide the extended ability-channel and application analyses supporting Section 4; they are summarised in the main text and retained here for completeness.

Table S1. Prominent and competing theoretical frameworks by ability channel (2020-2026), with representative implementing systems and key references.

Ability channel	Prominent post-2020 framework	Competing or complementary frameworks	Representative implementing systems	Key references
Perception	Predictive processing and active inference (generative, hierarchical percepts)	Marr/Tsotsos hierarchical vision; embodied active vision and affordances; Predictive Global Neuronal Workspace (PGNW)	CogNGen; CORTEX; MERLIN2; ArmarX	Ciria et al. (2021); Whyte and Smith (2021); Ororbia and Kelly (2022)
Attention	Attention Schema Theory plus precision-weighted active inference	Bayesian marginal-probability attention; transformer self-attention re-read as internal attention; salience gating	Attention-schema RL agents; information-sufficiency gates	Wilterson and Graziano (2021); Proietti et al. (2025)
Action selection	Expected-free-energy policy selection and dual-process arbitration	Soar operator selection with RL utility; BDI observe-decide-act; reactive/subsumption control	SOFAI; MAP; LLM-bootstrapped production agents	Fabiano et al. (2024); Webb et al. (2025); Proietti et al. (2025)
Memory	Complementary Learning Systems with episodic-memory desiderata for the LLM era	Hyperdimensional/vector symbolic memory; constructive reconstruction; global-workspace episodic access	EM-LLM; ArmarX memory; transformer-as-LTM; ACT-R/Soar stores	McClelland et al. (1995); Fountas et al. (2025); Wray et al. (2025)
Learning	Intrinsic-motivation/curiosity plus free-energy learning with LLM-supplied knowledge	Deep RL vs production learning; complementary one-shot vs gradual; analogical concept memory; lifelong open-ended learning	ACT-R pattern-discovery curiosity; CogNGen; LOLA robots	Modirshanechi et al. (2023); Ten et al. (2022); Nagashima et al. (2024)
Reasoning	Dual-process reasoning operationalized as neuro-symbolic arbitration	Analogical reasoning; non-monotonic constraint layers over LLM proposals; probabilistic-symbolic coupling	Shadow-production CMC bridges; verifier-gated LLM reasoners	Sun (2024); West et al. (2024); Acharya and Song (2025)

Ability channel	Prominent post-2020 framework	Competing or complementary frameworks	Representative implementing systems	Key references
Planning	Hierarchical/factored planning under dual-process control with modularized PFC functions	PDDL plus behaviour-tree/FSM stacks; hierarchical RL with modular reflection; tree/hypertree search	MAP; SOFAI; MERLIN2 (LLM plus classical planner)	Webb et al. (2025); Fabiano et al. (2024); González-Santamarta, Rodríguez-Lera, Fernández-Llamas, et al. (2024)
Metacognition	Nelson-Narens monitoring/control re-implemented as hybrid, hierarchical control	Higher-order self-model calibration; meta-d' signal-detection sensitivity; model-based metacognition	CMC-with-metacognition; intrinsic metacognitive learners	Laird, Lebiere, Rosenbloom, and Stocco (2026); Ukov and Tsochev (2025); T. Liu and van der Schaar (2025)
Communication	LLMs as the implicit-language competence layer under architectural control	LLM as declarative-memory oracle; LLM as rule/code generator; grounded multimodal chat managers	Sun's CLARION-LLM; MERLIN2 plus LLM; co-constructive HRI	Sun (2024); Scheibl et al. (2025); Allgeuer et al. (2024)
Motivation/emotion	Intrinsic motivation (curiosity, learning progress, empowerment) as core drive	Expected value of control; information-gap/compression-progress curiosity; affective modulation of modules	PILLAR-Robots; CogNGen; GWT-plus-valence agents	Modirshanechi et al. (2023); Ten et al. (2022); Nagashima et al. (2024)
Social cognition	Bayesian theory of mind over BDI, extended with abstraction	Baron-Cohen mindreading components on GWT substrate; ToM as intrinsic motivation; sequential/Bayesian ToM	TOMA; ACT-R machine ToM; CASPER	Z. Wang, Davis, et al. (2025); Erdoğan et al. (2025); Lebiere, Pirolli, et al. (2024)
Neuro-symbolic integration	Neuro-symbolic AI targeting robustness, uncertainty quantification, and intervenability	Logic tensor networks; CMC shadow-production bridging; graph-neural integration layers; IWMT synthesis	CogNGen; CMC/generative-network bridges; CogRec-style coupling	W. Wang and Yang (2024); Acharya and Song (2025); Safron (2020)

Table S2. Theoretical framework to implementing-architecture mapping, with core computational commitments and cross-ability reach.

Theoretical framework	Core computational commitment	Architectures/systems that instantiate it	Cross-ability reach	Key references
Predictive processing / active inference	Hierarchical generative model; prediction-error minimization; precision; expected free energy	CogNGen; CORTEX; IWMT; PGNW; active-inference robots	Perception, attention, action, learning, motivation	Ciria et al. (2021); Friston et al. (2017); Whyte and Smith (2021)
Attention Schema Theory	An accessible internal model of the agent's own attentional state used for control	Attention-schema RL agents	Attention, social coordination	Graziano and Webb (2015); Wilterson and Graziano (2021)
Complementary Learning Systems / episodic memory	Fast one-shot episodic capture plus slow semantic consolidation; cue-based reconstructive retrieval	EM-LLM; ArmarX; transformer-as-LTM; ACT-R/Soar memory	Memory, learning, metacognition	McClelland et al. (1995); Fountas et al. (2025)
Curiosity / intrinsic motivation	Novelty, information gain, learning progress, and empowerment as intrinsic reward signals	ACT-R pattern-discovery; CogNGen; LOLA; PILLAR-Robots	Learning, motivation, action	Oudeyer and Kaplan (2007); Ten et al. (2022); Modirshanechi et al. (2023)
Dual-process cognition	Fast associative proposer plus slow controlled verifier, with explicit arbitration	CLARION; SOFAI; MAP; CogniDual-style self-trained LLMs	Reasoning, planning, action, metacognition	Evans and Stanovich (2013); Sun (2024); Fabiano et al. (2024)
Common Model of Cognition (with metacognition)	Recurrent cycle over working, declarative, and procedural memory; self-representations for meta-control	Soar; ACT-R; Sigma; CogNGen; CMC-metacognition proposal	All twelve channels (specification-level)	Laird, Lebiere, and Rosenbloom (2017); Laird, Lebiere, Rosenbloom, and Stocco (2026); Ororbia and Kelly (2024); West et al. (2024)
Nelson-Narens metacognition	Object-level cognition monitored and controlled by an explicit meta level	CMC-with-metacognition; hierarchical metacognitive controllers; intrinsic metacognitive learners	Metacognition, reasoning, safety	Nelson and Narens (1990); Ukov and Tsochev (2025); T. Liu and van der Schaar (2025)

Theoretical framework	Core computational commitment	Architectures/systems that instantiate it	Cross-ability reach	Key references
Computational theory of mind (Bayesian/inverse planning)	Agent-indexed belief-desire models inferred by inverse planning, bounded by abstraction	TOMA; ACT-R machine ToM; CASPER; adaptive HRI	Social cognition, communication, planning	Jara-Ettinger (2019); Z. Wang, Davis, et al. (2025); Erdoğan et al. (2025)
Neuro-symbolic integration	Typed bidirectional exchange between learned and symbolic representations with provenance and uncertainty	CMC shadow-production bridges; CogNGen; logic-tensor and graph-neural couplings	Reasoning, perception, memory, integration	W. Wang and Yang (2024); Wan et al. (2025); Acharya and Song (2025)
Integrated World Modeling Theory (IWMT)	Fusion of free-energy/active inference with global-workspace and integrated-information accounts	IWMT reference model; active-inference consciousness architectures	Perception, integration, motivation	Safron (2020); Whyte and Smith (2021)

Table S3. Ability-by-ability synthesis of the post-2018 cognitive-architecture landscape.

Ability channel	2018-oriented baseline	Post-2018 development	Prominent frameworks and systems	Unresolved issue
Perception	Modality-specific pipelines feeding symbolic labels	Generative and multimodal perception is incorporated into hybrid control loops	Predictive processing/active inference; CogNGen; CORTEX; multimodal robot interfaces (Ciria et al. 2021; Ororbia and Kelly 2022; Trinidad Barnech et al. 2024)	Grounding and causal world-model validation under distribution shift
Attention	Salience and task-driven selection	Attention is increasingly represented as an internal control object rather than only a filter	Attention Schema Theory; precision weighting; information-sufficiency gates (Wilterson and Graziano 2021)	Benefits beyond constrained visuospatial tasks and separation from transformer attention
Action selection	Production-rule or utility-based arbitration	Fast learned proposals are increasingly governed by slower deliberative or metacognitive control	Dual-process arbitration; active inference; SOFAI; MAP (Fabiano et al. 2024; Webb et al. 2025)	When arbitration improves accuracy enough to justify latency and complexity

Ability channel	2018-oriented baseline	Post-2018 development	Prominent frameworks and systems	Unresolved issue
Memory	Working, declarative, procedural, and architecture-specific episodic stores	Episodic and multi-store memory becomes central to long-horizon agents and LLM augmentation	EM-LLM; ArmarX memory; ACT-R/Soar memory; memory-for-thoughts (Fountas et al. 2025; Peller-Konrad et al. 2023; Sievers and Russwinkel 2025)	Encoding specificity, deliberate cue construction, consolidation, forgetting, and provenance
Learning	Chunking, reinforcement learning, statistical association	Lifelong, intrinsically motivated, language-assisted, and analogical learning are combined	Production-model reinforcement learning; curiosity; free-energy learning; analogical concept memory; LLM knowledge acquisition (Brasoveanu and Dotlačil 2021; Modirshanechi et al. 2023; Mohan et al. 2020; Nagashima et al. 2024; A. Romero et al. 2023)	Longitudinal evidence for transfer without catastrophic forgetting
Reasoning	Symbolic inference layered on subsymbolic recognition	Neural or LLM proposals are verified, constrained, or explained by symbolic processes	Neuro-symbolic taxonomies; dual-process reasoning; Common Model bridges (W. Wang and Yang 2024; West et al. 2024)	Bidirectional symbol emergence and uncertainty propagation across the seam
Planning	HTN, PDDL, means-ends analysis	Hierarchical and factored planning combines LLM priors with explicit search and monitoring	SOFAI; MERLIN2; MAP; planning plus verification (Fabiano et al. 2024; González-Santamarta, Rodríguez-Lera, Fernández-Llamas, et al. 2024; Webb et al. 2025)	Generalization, recovery, and compute-accuracy trade-offs in long horizons

Ability channel	2018-oriented baseline	Post-2018 development	Prominent frameworks and systems	Unresolved issue
Metacognition	Optional or architecture-specific meta-reasoning	Self-models, confidence, memory for thoughts, and controller selection become explicit modules	CMC metacognition proposal; computational metacognitive architectures (Laird, Lebiere, Rosenbloom, and Stocco 2026; T. Liu and van der Schaar 2025; Ukov and Tsochev 2025)	Comparable measures of monitoring accuracy, control value, and self-repair
Communication	Scripted or grammar-based language modules	LLMs support open-ended dialogue, instruction parsing, and co-constructive learning under architectural control	MERLIN2+LLM; grounded multimodal HRI; co-constructive language architectures (Allgeuer et al. 2024; Scheibl et al. 2025)	When linguistic output can be trusted as evidence or executable intent
Motivation/ emotion	Fixed drives, appraisal, and homeostatic loops	Curiosity, learning progress, and empowerment become computational objectives and control signals	Curiosity taxonomies; ACT-R pattern-discovery motivation (Modirshanechi et al. 2023; Nagashima et al. 2024; Ten et al. 2022)	Multi-drive conflict, affective regulation, and safe exploration
Social cognition	Rule-based or limited theory-of-mind approximations	Bayesian and cognitively modeled theory of mind is integrated into HRI and personalization	CASPER; adaptive HRI; machine theory of mind (Freire et al. 2024; Lebiere, Pirolli, et al. 2024; Vinanzi and Cangelosi 2024)	Scalable social inference, recursive depth, norms, and evaluation with diverse users
Neuro-symbolic integration	Architecture-specific symbolic-subsymbolic combinations	Neuro-symbolic integration becomes an explicit design and research program	Robustness, uncertainty, intervenability, and compositional design (Acharya and Song 2025; Wan et al. 2025; W. Wang and Yang 2024)	Standardized interfaces, semantics, calibration, and systems efficiency

Table S4. Practical application domains and architecture-level value propositions.

Domain	Representative systems or lineages	Architecture-level value	Critical deployment requirement
Service and social robotics	MERLIN/MERLIN2, CASPER, ACT-R + Pepper, mind-inspired HRI	Grounded dialogue, explicit memory, behavior monitoring, social adaptation	Clarification, safe action grounding, user modeling, and recovery from communication failure
Industrial collaboration and manufacturing	DAC-HRC, Soar disassembly, LLM-ACTR	Reusable task knowledge, explainable decisions, human-robot coordination	Cycle-time and safety validation; transfer across products, workers, and sites
Cyber-physical systems and autonomous vehicles	Soar multi-UAV systems, cognitive digital-twin concepts	Persistent world state, hierarchical decisions, simulation-to-operation linkage	Real-time constraints, formal safety envelopes, and state synchronization
Healthcare and assistive systems	Rehabilitation and dietary-assistance agents; service robots	Personalized memory, transparent recommendations, long-term interaction	Clinical accountability, privacy, consent, calibrated uncertainty, and human escalation
Cyberdefense	Soar plus generative components	Explicit goals, attacker models, traceable response policies	Adversarial robustness, provenance, reversible action, and audit logging
Recommendation and personalization	CogRec; ACT-R declarative/procedural memory	Explanations grounded in a user model rather than only correlation	Bias monitoring, user control, memory correction, and non-manipulative adaptation
Education and cognitive simulation	LLM-assisted ACT-R/Soar model authoring	Lower modeling cost and testable cognitive hypotheses	Model validation, reproducible prompts, and separation of plausible text from cognitive evidence

C Statements and Declarations

Funding

The authors received no specific funding for this work.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval

Not applicable. This study synthesized published literature and did not involve human participants, animals, or identifiable private data.

Consent to Participate and Consent for Publication

Not applicable.

Received 12 July 2026; accepted 19 September 2026