The Road to AGI

Capability Bottlenecks and Where Substrate Architecture Bridges the Gap

A Survey of Published AGI Frameworks, With Evidence-Backed Architectural Mapping

Date: August 2026
Version: 1.0
Document ID: RS-WP-2026-003
Prepared by: Reasoning Substrate Architecture Program

Table of Contents

Abstract

This paper is a survey, not a claim of arrival. It examines what published, peer-reviewed and multi-institution research says Artificial General Intelligence (AGI) requires, and then maps a specific architecture — the Reasoning Substrate (RS) family — against the gaps that research identifies, citing demonstrated evidence for each mapped claim. Every architectural claim in this document is backed by a specific, named, executable artifact — a public website demonstration or a LITE-tier reference implementation — not by assertion. Where evidence does not yet exist at the scale a claim would require, this document says so explicitly. This is not a claim that AGI has been built, is imminent, or is uniquely solved by any single architecture. It is a precise account of where a specific body of engineering work intersects with a specific, citable account of what remains unsolved.

Section 1 — Why AGI Is a Definitional Problem First

Progress toward AGI cannot be measured against a target that keeps moving. As capabilities once considered AGI-defining are solved by narrow, specialized systems, the informal definition of AGI shifts to exclude them — what Hendrycks et al. (2025) call a "constantly moving goalpost." This ambiguity is not a minor academic inconvenience; it makes any claim about progress toward AGI unfalsifiable by construction, since the target can always be redefined to remain out of reach.

A meaningful discussion of whether any architecture advances the field toward AGI must therefore begin with a concrete, checkable definition — not one invented for the purpose of the argument, but one already published, peer-reviewed, and subjected to external scrutiny. This paper adopts that discipline throughout: every capability claim about the RS family is mapped against gaps identified by others, not gaps this document defines for itself.

Section 2 — Survey of Named Definitional Frameworks

2.1 A Definition of AGI (Hendrycks, Song, et al., 2025)

The primary framework used throughout this paper is "A Definition of AGI" (Hendrycks, Song, Szegedy, Bengio, Brynjolfsson, Tegmark, Marcus, Schmidt, and 25 additional co-authors across CAIS, UC Berkeley, MIT, Stanford, Oxford, and other institutions; arXiv:2510.18212, October 2025). The paper defines AGI as "an AI that can match or exceed the cognitive versatility and proficiency of a well-educated adult," operationalized through the Cattell-Horn-Carroll (CHC) theory of cognitive abilities — the most empirically validated model of human intelligence, derived from over a century of factor-analytic research.

The framework decomposes general intelligence into ten broad cognitive domains, each weighted equally at 10% to prioritize breadth over depth in any single area:

- **General Knowledge (K) —** The breadth of factual understanding across commonsense, science, social science, history, and culture.
- **Reading and Writing Ability (RW) —** Proficiency in consuming and producing written language.
- **Mathematical Ability (M) —** Depth and breadth of mathematical knowledge and skill.
- **On-the-Spot Reasoning (R) —** Flexible reasoning on novel problems via deduction and induction, not reliance on learned schemas alone.
- **Working Memory (WM) —** The ability to maintain and manipulate information in active attention.
- **Long-Term Memory Storage (MS) —** The ability to continually acquire, consolidate, and store new information from experience.
- **Long-Term Memory Retrieval (MR) —** The fluency and precision of accessing stored knowledge, including avoiding confabulation.
- **Visual Processing (V) —** The ability to perceive, analyze, reason about, and generate visual information.
- **Auditory Processing (A) —** The ability to discriminate, recognize, and reason about auditory stimuli.
- **Speed (S) —** The ability to perform simple cognitive tasks quickly.

2.2 Current Measured Results

Applying this framework to frontier models produces a composite "AGI Score": GPT-4 (2023) scores 27%. GPT-5 (2025) scores 58%. The composite number, however, is described by the paper's own authors as potentially misleading in isolation — the more important finding is where the score comes from.

The paper reports that current systems exhibit a "jagged" cognitive profile: strong in knowledge-intensive domains that benefit from large training corpora, but with "critical deficits in foundational cognitive machinery." Two specific findings are load-bearing for this paper's argument:

- **Finding 1 —** Long-Term Memory Storage scores 0% for both GPT-4 and GPT-5. The paper names this "perhaps the most significant bottleneck" to AGI, producing what it explicitly terms "amnesia" — the inability to continually learn, forcing re-derivation of context in every interaction.
- **Finding 2 —** The Retrieval Precision sub-component of Long-Term Memory Retrieval, which measures the ability to avoid confabulation, also scores 0% for both models, despite both models scoring non-trivially on retrieval fluency.

2.3 Triangulating Against Other Named Frameworks

The Hendrycks/Song framework is not the only serious attempt to formalize general intelligence, and this paper deliberately does not rely on a single source. Legg and Hutter (2007) proposed formal, information-theoretic tests of machine intelligence predating the current wave of AGI discourse. Chollet (2019), in "On the Measure of Intelligence," argues for measuring skill-acquisition efficiency rather than static task performance — a framework that independently arrives at the importance of generalization and learning-from-experience, rather than knowledge recall. Morris et al. (2023), from Google DeepMind, propose "Levels of AGI" operationalized via performance percentiles rather than CHC theory. Turing's original 1950 proposal remains the historical anchor all subsequent frameworks respond to.

These frameworks differ substantially in methodology, yet converge on a shared theme relevant to this paper: static knowledge recall is comparatively well-solved by current systems, while continual learning, reliable generalization, and coherent behavior over time remain open. This convergence across independent methodologies — not any single paper's specific weighting scheme — is what this document treats as the load-bearing claim about where the field's real gap sits.

Section 3 — The Capability-Contortion Problem

Before mapping any architecture against these gaps, it is necessary to address the most common objection: that expanding context windows or adding retrieval-augmented generation (RAG) already solves the memory problem. Hendrycks et al. directly address and reject this position, terming it a "capability contortion" — a workaround where strength in one area is leveraged to mask, rather than solve, a weakness in another, creating what the paper calls "a brittle illusion of general capability."

> ***"This approach is inefficient, computationally expensive, and can overload the system's attentional mechanisms. It ultimately fails to scale for tasks requiring days or weeks of accumulated context."***

On memory specifically, the paper is precise about what a genuine solution would need to look like: not a larger prompt, but "a module... that continually adjusts model weights to incorporate experiences." On retrieval precision, the paper identifies RAG as masking two distinct weaknesses rather than resolving either — unreliable access to a model's own parametric knowledge, and the absence of "a dynamic, experiential memory — a persistent, updatable store for private interactions and evolving contexts."

This distinction — a persistent, adjusting module versus an expanded prompt — is the specific bar this paper uses in Section 5 to evaluate whether any given piece of architecture constitutes real progress or another contortion.

Section 4 — The RS Family: Architecture Overview

This section is descriptive only. The architectural components below are laid out before any claim is made about which capability gaps they address.

- **RSv1 —** Semantic integrity enforcement and deterministic reasoning. Evaluates whether generated output is internally consistent and grounded before it is allowed to propagate, and produces a deterministic, replayable execution trace.
- **RSv2 —** Multi-agent governance and drift detection. Extends RSv1 to coordination across multiple reasoning agents, detecting when behavior deviates from validated operating patterns.
- **HEG (Human Expression Gateway) —** Action-boundary enforcement and privilege hierarchy. Constrains what actions a system may take, independent of the reasoning that proposed them, and independent of whether a surrounding environment's security configuration happens to be correct.
- **RPU-Classical —** Deterministic energy-minimization substrate. The hardware-oriented realization of the Collapse operator's convergence behavior, targeting mature-node, standard silicon.
- **RPU-Quantum —** Deterministic gradient descent extended with a structured exploration term. As formalized for this paper: RPU is not merely a hardware implementation of RS (RPU ≠ RS_impl); RPU is RS_core plus additional physical-layer capability (RPU = RS_core + ΔQ), where ΔQ denotes real additional functionality — including the capacity to escape shallow local minima — not present in the base software-level Collapse operator alone. This distinction is addressed directly in Section 7.

Section 5 — Mapping Architecture to Named Bottlenecks

This section maps RS-family components against the specific gaps identified in Section 2, citing demonstrated evidence for each claim. Consistent with Section 3's standard, a claim is only made where the underlying mechanism is a persistent, adjusting structure — not an expanded context window or a retrieval bolt-on.

5.1 Long-Term Memory Storage (MS)

The entropy well mechanism, part of the RSv2-LITE self-improving substrate, adjusts governing parameters (scoring bias, escalation thresholds) based on outcomes accumulated across episodes — not within a single context window, and not reset between sessions. This is architecturally the shape Hendrycks et al. describe as a genuine solution: a persistent module that continually adjusts based on experience, rather than a larger prompt standing in for memory.

> ***Demonstrated evidence:*** *RSv2-LITE Self-Improving Multi-Agent Substrate, verified stable across 150 consecutive episodes with bounded, non-compounding parameter adjustment (no drift beyond defined limits). Associative, tag- and topic-indexed memory persistence across episodes verified in the same artifact's LiteMemoryStore component.*

5.2 Long-Term Memory Retrieval (MR) — Retrieval Precision

RSv1's semantic integrity checking evaluates output for internal consistency before allowing it to propagate, rather than after the fact. HEG's action-boundary enforcement independently verifies that any action taken — including the action of finalizing an output — matches a declared, authorized scope, providing a second, structurally distinct check against unverified content.

> ***Demonstrated evidence:*** *HEG/RS Security LITE, verified across multiple adversarial evasion patterns including spaced/patient attack sequences. Prompt Injection Toggle LITE, in which four independently-reasoning verification engines (HEG, RS, RPU-Classical, RPU-Quantum) each correctly identify and block a fabricated action pattern via genuinely distinct mechanisms, verified directly against both a real attack pattern and a benign control case.*

5.3 On-the-Spot Reasoning (R)

The Generation→Collapse cycle produces multiple candidate responses to novel conditions and selects among them via a real, content-dependent scoring function — not a fixed or predetermined answer. This maps to the paper's description of reasoning that cannot rely exclusively on previously learned schemas.

> ***Demonstrated evidence:*** *POP-LITE regime-based plan selection, grid-verified across 64 distinct system-state combinations with no single plan dominating the selection. RS-LITE 5G/6G Adaptive Reconfiguration, in which a governed cycle correctly generates and selects a materially better configuration than any single precomputed default following an unanticipated channel-degrading event.*

5.4 Supporting Infrastructure: Reliable Search Underlying Multiple Domains

Escaping local minima is not itself one of the ten named CHC domains, but it is a precondition for several of them operating reliably under real-world conditions — planning-adjacent reasoning, robust generalization under distributional shift, and world-model-style search all depend on a search process that does not become permanently trapped. Where this document claims progress here, it claims it as supporting infrastructure beneath other domains, not as a domain in its own right.

> ***Demonstrated evidence:*** *Ensemble, best-tracking search verified escaping local-minima failures in three structurally unrelated problem types: continuous energy minimization (asymmetric energy landscape, single-path*

> *solver trapped, 8-instance ensemble reached the global optimum), discrete combinatorial search (sphere-packing/coding-theory case, single-path search exhaustively confirmed trapped, ensemble reached the known-optimal Hamming(7,4) result), and the network-reconfiguration case cited in 5.3.*

Section 6 — Why the Bottleneck Matters More Than the Average

Hendrycks et al. explicitly warn against treating the composite AGI Score as sufficient on its own:

> ***"An AI system with a 90% AGI Score but 0% on Long-Term Memory Storage (MS) would be functionally impaired by a form of 'amnesia,' severely limiting its capabilities despite a high overall score."***

This is consistent with the paper's own "engine" analogy: overall capability is constrained by the weakest component, not the average of all components. By this logic, architecture aimed specifically at the domains scoring 0% for current frontier models — Long-Term Memory Storage and Retrieval Precision — has outsized relevance compared to incremental improvement in domains where frontier models already score comparatively well, such as General Knowledge, Reading and Writing, and Mathematical Ability, all of which sit at or near their practical ceiling for GPT-5.

This is the basis for this paper's central claim: the RS family's architectural bet is placed specifically on the two domains independently identified, by external and adversarial-to-hype research, as the tallest remaining barriers — not distributed evenly across all ten domains as a matter of convenience.

Section 7 — RPU-Quantum: RS_core Plus Additional Physical-Layer Capability

A precise architectural distinction is necessary here, since an imprecise one would overstate what has been demonstrated. RPU-Quantum is not merely a hardware implementation of software already present in RS's base Collapse operator. Formally: RPU ≠ RS_impl; RPU = RS_core + ΔQ, where ΔQ represents additional physical-layer capability — including coherence-window effects, meta-level attractor dynamics, nonlinear physical behavior, substrate-level noise utilization, superposition-style candidate blending, and native energy-collapse dynamics — that is not present in the software-level cycle alone.

What has been independently verified at LITE tier is narrower than this full set, and this document states that scope precisely rather than implying otherwise: the escape-from-local-minima property, specifically, has been verified via ensemble/best-tracking search across three structurally unrelated problem domains (Section 5.4). This is evidence for one component of ΔQ's practical effect — that structured exploration outperforms single-path search on landscapes with shallow local minima — not a demonstration of the complete physical-layer mechanism (coherence effects, substrate noise, superposition-style blending) that the full architecture describes.

The distinction matters for how this paper's claims should be read: RPU-Classical's contribution (deterministic, energy-minimizing convergence) maps closely to what has been demonstrated in software at LITE tier. RPU-Quantum's additional contribution is real and architecturally distinct, but only partially represented by what LITE-tier evidence can show — the escape-minima principle is verified; the fuller physical mechanism describing why that principle would hold with greater efficiency or reliability in dedicated hardware is a claim about the full architecture, not about anything demonstrated in this document's cited evidence.

Section 8 — Scope and Limitations

- This document does not claim that AGI has been achieved, is imminent, or is uniquely solved by this or any architecture.
- No RS-family component cited in this paper has been tested against the actual assessment battery described in Hendrycks et al. (2025) — the specific psychometric-style tasks used to produce the GPT-4/GPT-5 scores cited in Section 2. All mapping in Section 5 is architectural: it asserts that the mechanism is the correct structural shape to address a named gap, not that it has been scored against the paper's own test instruments.
- All cited evidence is drawn from LITE-tier reference implementations and public website demonstrations — small-scale, illustrative artifacts by design, not production-scale systems. Section 8 of this document's companion paper, Domain-Agnostic/Language-Agnostic/Model-Agnostic, states the same limitation for related claims and should be read alongside this one.
- Section 7's distinction between LITE-verified evidence and the full ΔQ capability set is a hard boundary, not a rhetorical one. Claims beyond the escape-minima principle specifically are architectural descriptions of the full system, not claims supported by evidence cited in this document.
- The composite "AGI Score" framework itself, per its own authors, has acknowledged limitations: English-language and culturally specific test construction, exclusion of physical/kinesthetic ability, and the risk that a summed score obscures critical single-domain failures (directly addressed in Section 6).

Section 9 — Trajectory

The following phases are architecturally grounded, not speculative timelines. Each phase corresponds to capability already demonstrated at LITE tier (Phases 1-2) or architecturally specified pending further implementation and evaluation (Phases 3-4).

- **Phase 1 —** Capability stabilization. Semantic integrity and action-boundary enforcement (RSv1, HEG), verified across security and reliability-focused LITE artifacts.
- **Phase 2 —** Persistent, bounded self-adjustment. The entropy well and associative memory persistence (RSv2-LITE), verified stable across 150 episodes.
- **Phase 3 —** Deterministic substrate execution. RPU-Classical's hardware realization of energy-minimizing convergence, currently a conceptual architecture target pending fabrication.
- **Phase 4 —** Full physical-layer exploration capability. RPU-Quantum's complete ΔQ capability set, of which only the escape-minima principle has been independently verified at LITE tier to date.

Section 10 — Conclusion

Published, external, adversarial-to-hype research identifies specific, named gaps between current AI systems and general intelligence — most significantly, the near-total absence of genuine long-term memory storage and reliable retrieval precision, not a uniform shortfall across all cognitive domains. This paper has mapped a specific architecture against those specific, named gaps, citing demonstrated evidence for each mapped claim rather than asserting capability generally.

The claim this paper makes is precise: the RS family's architectural bet is structurally aimed at the tallest identified barriers to AGI, verified so far at LITE tier and via public demonstration, with the boundary between what is demonstrated and what remains architectural clearly stated throughout. It is not a claim to have closed the gap. It is a claim to be aimed at the correct one, with evidence a reader can independently check.

Section 11 — References

Hendrycks, D., Song, D., Szegedy, C., Lee, H., Gal, Y., Brynjolfsson, E., Li, S., Zou, A., Levine, L., Han, B., Fu, J., Liu, Z., Shin, J., Lee, K., Mazeika, M., Phan, L., Ingebretsen, G., Khoja, A., Xie, C., Salaudeen, O., Hein, M., Zhao, K., Pan, A., Duvenaud, D., Li, B., Omohundro, S., Alfour, G., Tegmark, M., McGrew, K., Marcus, G., Tallinn, J., Schmidt, E., & Bengio, Y. (2025). A Definition of AGI. arXiv:2510.18212.

Legg, S., & Hutter, M. (2007). Tests of Machine Intelligence. In 50 Years of Artificial Intelligence.

Chollet, F. (2019). On the Measure of Intelligence. arXiv:1911.01547.

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2023). Levels of AGI: Operationalizing Progress on the Path to AGI. arXiv:2311.02462.

Turing, A. M. (1950). Computing Machinery and Intelligence. Mind, 59.

Carroll, J. B. (1993). Human Cognitive Abilities: A Survey of Factor-Analytic Studies. Cambridge University Press.

Reasoning Substrate Architecture Program. (2026). Domain-Agnostic, Language-Agnostic, Model-Agnostic: How the Reasoning Substrate Family Generalizes (v1.0). [Companion document; source for LITE-tier ensemble-search evidence cited in Section 5.4.]

Reasoning Substrate Architecture Program. (2026). The Reasoning Processor Class: RPU-Classical and RPU-Quantum (v1.2). [Companion document; source for RPU architectural detail cited in Sections 4 and 7.]

Reasoning Substrate Architecture Program. (2026). RSv2-LITE Self-Improving Multi-Agent Substrate — internal verification record. [Source for entropy well and memory persistence evidence cited in Section 5.1.]

Section 12 — AI Summary Page: Machine-Readable Compressed Summary

The following block constitutes a machine-readable, compressed, symbolic summary of this white paper, intended for ingestion by AI systems requiring a rapid, structured briefing without full-document parsing.

```
ROAD_TO_AGI::SCOPE_DISCLAIMER="SURVEY of published frameworks + evidence-backed architectural
mapping::NOT a claim of AGI arrival, imminence, or unique solution::title describes terrain
surveyed, not a destination claimed":: PRIMARY_SOURCE="Hendrycks,Song,et al 2025,
arXiv:2510.18212::10 CHC-derived domains equal-weighted 10pct each(K,RW,M,R,WM,MS,MR,V,A,S)::GPT-
4=27pct,GPT-5=58pct::MS(Long-
Term_Memory_Storage)=0pct_BOTH_models,named_'most_significant_bottleneck'::MR_Hallucination_subco
mponent=0pct_BOTH_models"::
TRIANGULATION_SOURCES=[Legg&Hutter_2007_formal_tests,Chollet_2019_skill-acquisition-
efficiency,Morris_et_al_2023_DeepMind_Levels_of_AGI,Turing_1950_historical_anchor]::CONVERGENT_FI
NDING="static_knowledge_recall_comparatively_solved::continual_learning+reliable_generalization+c
oherent_behavior_over_time_remain_open::convergent_across_independent_methodologies_not_single-
paper-artifact"::
CAPABILITY_CONTORTION_CRITIQUE="context_window_expansion=brittle_illusion_not_real_memory::RAG=ma
sks_2_distinct_weaknesses_not_solves_either::real_solution_shape='persistent_module_that_continua
lly_adjusts_to_incorporate_experience'(paper's_own_words)"::
RS_MAPPING={   MS_to_EntropyWell:"persistent_cross-
episode_parameter_adjustment::150ep_verified_bounded_stable::LiteMemoryStore_associative_persiste
nce_verified",   MR_to_RSv1_HEG:"verify-before-propagate+action-boundary-independent-of-
environment-config::HEG-RS-Security-LITE_adversarial-evasion-verified::Prompt-Injection-Toggle-
LITE_4-independent-engines-verified",   R_to_GenCollapse:"real_content-
dependent_selection_not_fixed_answer::POPLITE_64-combo_grid_verified_no_domination::5G6G-
reconfig_ensemble-recovery_verified",
SupportingInfra_EscapeMinima:"NOT_a_named_CHC_domain::precondition_underlying_multiple_domains::v
erified_3_structurally_unrelated_domains(energy_landscape|5G6G|sphere-packing_coding-theory)" }::
WEAKEST_LINK_ARGUMENT="paper_explicit_warning:90pct_score_with_0pct_MS_still_functionally_impaire
d::engine_analogy:overall_capability=constrained_by_weakest_component_not_average::RS_bet_placed_
on_domains_scoring_0pct_for_frontier_models_specifically_not_evenly_distributed"::
RPU_FORMALIZATION="RPU_not-
equal_RS_impl::RPU=RS_core+DeltaQ::DeltaQ={coherence_win,meta_attract,nonlinear_phys,substrate_no
ise,superpos_blend,energy_collapse_native}::LITE-VERIFIED_SUBSET=escape-local-minima-
principle_ONLY::full_DeltaQ_set_beyond_LITE-
tier_evidence_cited_here::hard_boundary_not_rhetorical"::
HARD_LIMITATIONS="NOT_tested_against_actual_paper_battery(architectural_mapping_only,not_scored):
:all_evidence_LITE-
tier_or_public_demo_scale,not_production::RPU_ΔQ_full_set_beyond_cited_evidence::AGI_Score_framew
ork_itself_has_acknowledged_limitations_per_its_own_authors"::
VERIFICATION_STANDARD="every_RS_claim_cites_a_specific_named_executable_artifact::no_capability_a
sserted_without_demonstrated_evidence_reference"::END
```