November 7, 2025

AGITB: A Signal-Level Benchmark for Evaluating Artificial General Intelligence

Matej Šprogar MATEJ.SPROGAR@UM.SI
Faculty of Electrical Engineering and Computer Science
University of Maribor
SI-2000 Maribor, Slovenia

Editor:

Abstract

While current artificial intelligence systems demonstrate remarkable capabilities, they remain specialized and lack a unified measure of general intelligence. Existing evaluation frameworks, which focus primarily on language or perception tasks, offer limited insight into generality. The Artificial General Intelligence Testbed (AGITB) introduces a complementary benchmarking suite comprising fourteen elementary tests, thirteen of which are fully automated.

AGITB evaluates models on their ability to forecast the next input in a temporal sequence, step by step, without pretraining, symbolic manipulation, or semantic grounding. The framework isolates core computational invariants, such as determinism, sensitivity, and generalisation, that parallel principles of biological information processing. Designed to resist brute-force or memorisation-based strategies, AGITB enforces unbiased and autonomous learning. The human cortex satisfies all tests, whereas no current AI system meets the full AGITB criteria, demonstrating its value as a rigorous, interpretable, and actionable benchmark for evaluating progress toward artificial general intelligence. A reference implementation of AGITB is freely available on GitHub.

Keywords: artificial general intelligence, benchmarking, generalisation, symbol grounding problem, temporal sequence prediction

1 Introduction

Despite rapid advances in machine learning and neural network architectures, artificial intelligence (AI) systems still lack the flexibility and robustness of human intelligence. Marcus and Davis (2020) observed that although large language models (LLMs) can generate highly fluent outputs, they rely primarily on statistical pattern recognition rather than grounded, compositional reasoning. Mitchell (2025) further summarised why even the latest large reasoning models (LRMs), despite some impressive achievements, cannot be trusted. Surface-level competence obscures a deeper lack of understanding, which is an essential prerequisite for artificial general intelligence (AGI).

The progress in AI has led to a growing speculation that AGI is near. However, such claims remain difficult to substantiate without a rigorous and informative metric. Assessing progress toward AGI requires more than specialised metrics or qualitative impressions; it necessitates principled, general-purpose metrics capable of systematically capturing and comparing essential cognitive capabilities.

Various attempts have been made to define such tools, the most iconic being the Turing test proposed by Turing (1950). However, no existing test fully achieves its intended purpose; moreover, all lack key properties such as gradual resolution, interpretability, and full automatizability. The metrics typically assess surface-level competence rather than underlying mechanisms of generalisation. In response to these limitations, this paper introduces the Artificial General Intelligence Testbed (AGITB), a novel benchmark for validating foundational cognitive abilities in artificial intelligence systems.

AGITB draws inspiration from the Ladder to Human-Comparable Intelligence introduced by Šprogar (2018), extending its conceptual foundations into a unified testbed comprising fourteen core requirements, thirteen of which are implemented as fully automatable and domain-agnostic tests. Although AGITB does not aim to evaluate consciousness or semantic comprehension, it provides a principled framework for distinguishing narrow AI systems from those exhibiting generalisable, adaptive behaviour. To position AGITB within the broader landscape of AGI evaluation, we include a comparative analysis with the Abstraction and Reasoning Corpus (ARC; Chollet, 2019) and the NeuroBench framework (Yik et al., 2025).

2 Background

The rapid progress of deep learning has enabled AI systems to exhibit increasingly sophisticated reasoning, problem-solving, and dialogue capabilities. However, despite these advances, a persistent reluctance remains to attribute "intelligence" to machines. This hesitation is partly rooted in the intuitive association of intelligence with uniquely human traits, such as consciousness, self-awareness, and subjective experience, which remain elusive in artificial systems.

Historically, as AI systems have succeeded in domains once considered hallmarks of human intelligence, definitions of intelligence have undergone significant shifts. For instance, the success of Deep Blue against Garry Kasparov in chess (a task formerly seen as a benchmark for AGI) was quickly reframed as a triumph of brute-force computation rather than genuine intelligence. Such redefinitions risk obscuring real milestones. As AI approaches human-level capabilities, we may inadvertently set a perpetually receding goalpost for AGI, failing to recognise it even when achieved.

Although AGI is typically envisioned as matching human cognitive flexibility across diverse domains, its evaluation has largely defaulted to narrow, task-specific metrics. This is partly due to the absence of a universally accepted AGI benchmark. Researchers have thus gravitated toward achieving superhuman performance in discrete domains, where progress can be clearly quantified. However, such specialised benchmarks favour narrow AI by rewarding depth within isolated subdomains rather than breadth of adaptation and general reasoning—hallmarks of general intelligence. Ironically, some of these benchmarks are now so specialised that humans have difficulty with them.

2.1 A Benchmark That Only Humans and AGI Can Meet

An effective AGI test must be trivial to solve for humans yet remain inaccessible to contemporary machine learning models that rely on brute-force methods, pretraining, or statistical

pattern matching. Such a test must demand capabilities that transcend memorisation or domain-specific heuristics, requiring generalisation, abstraction, and adaptive reasoning.

A valid AGI evaluation must either (1) reveal and exploit a fundamental cognitive gap between humans and machines or (2) define a behavioural capacity that current artificial systems cannot replicate. The first strategy is increasingly fragile, as advanced AI systems often bypass genuine understanding through massive pretraining on diverse datasets. The second strategy may involve a more faithful emulation of human cortical computation, pointing to neuromorphic architectures such as spiking neural networks, which more closely mirror the brain's time-sensitive, event-driven dynamics, as described by Maass (1997).

In alignment with this biologically grounded perspective, the AGITB departs from symbolic, language-based evaluations and instead assesses intelligence at the lowest, signal-processing level. While Turing was right to suggest that communication could serve as a basis for evaluating machine intelligence, natural language remains problematic as a test medium: it depends on shared human experiences and symbols whose meanings are ungrounded in machines, as argued by Harnad (1990).

AGITB thus adopts a more elemental approach. Rather than judging intelligence by symbolic interpretation, it evaluates whether a system can detect, learn, and generalise patterns in raw binary signals. A neural spike by itself contains the smallest amount of information possible and is, as such, grounded but free of other semantics. A binary signal accurately represents the neural spike.

Building on the view of Hawkins and Blakeslee (2004) that intelligence is fundamentally about extracting structure from data to enable prediction, AGITB operates at the level of signal-based forecasting. This approach aligns closely with the functioning of biological intelligence at the cortical level, which processes time-sensitive sensory spike trains rather than disembodied symbols.

3 Artificial General Intelligence Testbed

The testbed supports the development and evaluation of AGI by defining a clear set of requirements that a model under evaluation must meet to qualify as generally intelligent. A model must satisfy *all* requirements in order to claim success on the benchmark.

The central premise is that the capacity to pass the AGITB constitutes a necessary condition for the existence of AGI. Non-AGI systems cannot meet AGITB's requirements. Such a premise cannot be proven; it can only be rejected by any non-AGI system that meets the AGITB criteria. The fact that humans can pass the AGITB suggests that AGITB requires a solution beyond the current state-of-the-art AI.

Unfortunately, without a clear definition of intelligence, we can only speculate that AGITB defines the necessary (but not sufficient) building blocks for the emergence of genuine, domain-independent intelligence. This assumption of necessity stems from the fact that, except for human brains, no existing artificial system has met its requirements.

However, an AGITB-passing system is not automatically also a fully fledged AGI; after all, it is not required to express higher-level skills, like reasoning or natural language fluency. However, a failure to meet the AGITB's requirements likely indicates a deficiency in foundational general cognitive capabilities.

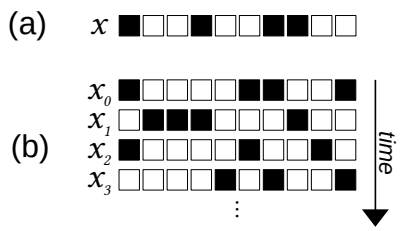

Figure 1: (a) Example of a 10-bit input x with 4 bits set.
(b) Example of an input sequence showing a temporal pattern of period four.

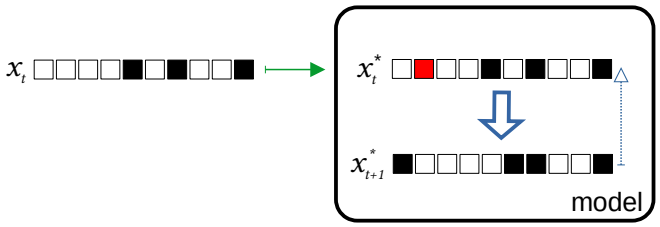

Figure 2: Iterative adaptation in discrete time. At the previous time step $(t-1)$, the model issued a prediction x_t^*, with the second bit incorrect in this instance. After observing the realized input x_t, the model updates and subsequently produces a correct one-step-ahead prediction x_{t+1}^* for the sequence shown in Figure 1b.

3.1 Components

AGITB evaluates the user-supplied AGI model as a black box that predicts future inputs from the historical sequence of observed signals. Each input comprises ten bits, the specific semantics of which are immaterial; each bit may correspond to an arbitrary channel, such as a pixel, an audio band, or actuator feedback.

Each input represents a snapshot of multiple parallel signals at a single time step (Figure 1a). Spatial organisation within each input encodes local structure, whereas semantic richness arises from the temporal evolution of the input sequence (Figure 1b). The interaction between spatial and temporal dimensions gives rise to structured patterns that are challenging for the model to adapt to.

3.2 Operation

The testbed presents the model with a stream of inputs over time. At each time step t, the model receives an input p_t and is required to predict the subsequent input p_{t+1}, as shown in Figure 2. The central challenge is not simply extrapolation but discerning the underlying causes or regularities that produce the observed input stream and using that understanding to make accurate future predictions.

3.2.1 Design

A distinguishing architectural feature of AGITB is its deliberate avoidance of conventional correctness metrics such as accuracy or mean-squared error. The limitations of these metrics are twofold. First, they cannot reliably distinguish between AGI and human performance, or that of non-AGI systems, as contemporary AI models can already surpass humans on standard benchmarks. Second, in an era of elevated expectations driven by specialised generative AI systems, the anticipated AGI performance on such metrics is often set unrealistically high. Even well-educated humans possessing fully developed and highly parallelised brains may underperform relative to current AI systems. It is therefore unreasonable to expect a first-generation AGI, potentially operating on a simplified and computationally constrained simulation of the brain, to match or exceed human-level results. Currently, our understanding of neural mechanisms is insufficient to justify such expectations.

Without conventional metrics or predefined performance thresholds, it can be difficult to determine whether a model has achieved a meaningful level of competence. AGITB addresses this challenge by employing a self-referential evaluation approach, in which the model under test is compared against its own independent instances. Each test constructs a controlled scenario involving one or more such models, whose internal states and behaviours are analysed comparatively. Requiring each test to be passed 100 times renders AGITB an extreme form of stress testing, ensuring that successful performance reflects genuine robustness rather than chance. Success is thus defined in terms of the relative consistency or superiority of model responses, rather than by any external quantitative metric.

For these reasons, and to minimise type I errors, AGITB employs an all-or-nothing criterion: the system under evaluation must successfully pass *all* fourteen tests. This design choice is justified, as individual tests are solvable by non-AGI systems, whereas the simultaneous satisfaction of all fourteen requirements is likely to demand capabilities beyond narrow intelligence. AGITB therefore posits that what non-AGI systems lack is the property referred to as "intelligence," understood as the synergistic integration of all fourteen tests.

3.3 The 14 AGI Requirements

The reference implementation of AGITB in C++ is freely available under the GPL-3 license: `https://github.com/matejsprogar/agitb`.

Requirement 1 – Unbiased start

Assertion: Models that have received no input are considered empty and thus equal.
Assertion: An empty model predicts an empty input.

AGITB rests on the foundational assumption that general-purpose learning systems, such as the brain, do not possess an innate understanding of external inputs. Instead, they learn through interaction with their environment. Each system must independently construct meaning from raw sensory data without relying on pre-encoded semantics. While some reflexes may be genetically predetermined, these do not constitute genuine understanding. In this respect, all cortical systems begin from an unbiased initial state, shaped entirely by the input received over time.

A model that predicts non-empty (non-zero) inputs prior to receiving data is considered biased, as it encodes unwarranted assumptions about the future. To ensure neutrality, models must initialise with empty predictions and remain unbiased until conditioned on actual input.

Two model instances that have received no input should compare as equal, as neither has been influenced by prior experience. Here, "empty" refers to a lack of informational content, not structural capability. The model must possess an intrinsic organisational architecture capable of learning, even in the absence of prior data.

The test operationalises, in empirical form, the philosophical challenge posed by classical symbolic systems: the emergence of meaning without prior grounding.

Requirement 2 – Bias

Assertion: Any model that has received input can no longer be regarded as unbiased.

Every input modifies the model's state, shaping its internal dynamics. As such, the mere act of input processing introduces bias based on experience.

Requirement 3 – Determinism

Assertion: If two cortices are equal, they must have received the same inputs.

Biological neurons operate in a functionally deterministic manner, ensuring stability and consistency in brain function. Although small stochastic effects may occur, they do not undermine the structured, rule-based nature of neural processing. Similarly, in AGITB, two functionally identical models must have experienced identical input histories. Any difference in input must lead to a divergence in state, reinforcing the principle that behaviour and internal state are fully determined by input alone. The brain's actions are effectively deterministic, though often perceived otherwise due to their immense complexity, as noted by Cave (2016) in his discussion of free will.

Determinism at the level of neural signal processing remains a necessary condition for the stable and reproducible functioning of the brain, even if its manifestations at the cognitive level are obscured by complexity and individual variation.

Requirement 4 – Sensitivity

Assertion: Two different cortices remain different, even if they experience long exposure to identical inputs.

Cortical systems exhibit chaotic sensitivity to initial conditions. Small differences in early experiences or internal states can lead to divergent trajectories over time. This deterministic sensitivity, amplified through complex interactions, accounts for the illusion of unpredictability in decision-making.

Requirement 5 – Time

Assertion: Changing the input order results in a different model state.

Because the model updates its state based on cumulative history, the order of inputs critically shapes learning and adaptation. The ability to recognise and exploit temporal structure is therefore a defining property of intelligent systems.

However, sensitivity to input order alone is not sufficient. Since temporal coding makes time an intrinsic property of neural data, Requirement 6 is designed to rule out models that perform equally well on temporally invariant data, as such models fail to exploit temporal information.

Requirement 6 – Refractory period

Assertion: The model must be able to adapt to any minimal-period input sequence that respects proper refractory periods.

Assertion: The model cannot adapt to an input sequence that repeats a neural spike in violation of refractory-period constraints.

Biological intelligence depends on discrete spikes for signal transmission and learning. AGITB enforces refractory periods to reflect this constraint, capturing the biological principle that a neuron cannot fire again *immediately* after activation.

While refractory periods are not the source of spiking variability, they impose a minimum separation between spikes, preventing continuous or unmodulated firing. This constraint helps preserve the temporal diversity needed for synaptic adaptation. Gerstner and Kistler (2002) showed that input sequences that lack sufficient variability, such as constant or overly repetitive spiking, fail to support effective learning. Consequently, AGITB permits only those temporal sequences that respect biologically plausible refractory dynamics while remaining agnostic to any particular semantic encoding of signals.

The refractory period requirement does not preclude the possibility of more general solutions. A universal system might simulate compliance with this constraint without relying on genuine refractory dynamics, whereas a system based on refractory processing cannot simulate the absence of such dynamics.

Requirement 7 – Temporal adaptability

Assertion: The model must be able to adapt to a temporal pattern with the user-specified period.

Assertion: The model must also be capable of adapting to a temporal pattern with a period longer than the user-specified value.

This test assesses the model's ability to learn across different temporal scales. Unlike rigid temporal pattern-matching systems, intelligent models should be able to detect and respond to recurring structures, regardless of their exact periodicity. The purpose of this test is to exclude models that cannot adapt to temporal patterns of different lengths.

Requirement 8 – Stagnation

Assertion: There exists a limit beyond which the model can no longer adapt, even to patterns that would otherwise be learnable.

Cognitive systems inevitably reach adaptation limits as their finite resources become saturated. This test assesses whether such limits emerge over time when the model is exposed to input sequences that would otherwise be learnable.

Requirement 9 – Content sensitivity

Assertion: Adaptation time depends on the content of the input sequence.

The structure and regularity of an input sequence influence the rate at which a model adapts, where adaptation time is defined as the number of iterations required to predict the entire temporal pattern accurately. Simpler or more predictable sequences typically result in faster convergence, while irregular or noisy inputs require greater exposure before the model can reliably learn and reproduce them.

Requirement 10 – Context sensitivity

Assertion: Adaptation time depends on the state of the model.

Past learning influences how new information is integrated. A model with a well-structured internal state may adapt quickly to familiar or related patterns, whereas unstructured or conflicting states may require more extensive reorganisation.

Requirement 11 – Unobservability

Assertion: Distinct cortices may exhibit the same observable behaviour.

Identical external outputs can arise from different internal states. This many-to-one mapping underscores that observable behaviour alone cannot reveal the underlying structure or history of a model's internal dynamics.

Requirement 12 – Denoising

Assertion: A model can recall a sequence despite perturbations.

An intelligent model should be capable of efficiently memorising input patterns in a manner that enables the recall of previously observed sequences even when these are perturbed by noise. When re-exposed to familiar stimuli, such a model is expected, on average, to outperform random guessing in predictive tasks. Average performance is used as the evaluation criterion, since random models may occasionally produce correct predictions by chance without reflecting genuine learning or structural understanding.

Test 12 is designed as a deliberately conservative benchmark. The model is required to outperform random guessing in all 100 independent trials. This procedure is not interpreted as a conventional significance test (e.g., at the 5% level); rather, it serves as a stringent robustness check. The design ensures that only large and systematic performance gains result in a pass. Modest or marginal improvements, while potentially real, are intentionally treated as failures, as the objective is to identify only clear and substantial advances in model capability.

Requirement 13 – Generalization

Assertion: The model performs above chance on previously unseen inputs.

Only models capable of generalisation can derive lasting benefit from prior learning. After exposure to a given set of stimuli, such models are expected, on average, to outperform random baselines when predicting future inputs. Average performance is used as the evaluation criterion, since even a random model may occasionally produce correct predictions by chance without demonstrating genuine understanding or the acquisition of structured knowledge.

The generalisation assessment follows the same conservative design as Test 12. By requiring success across all repetitions, the procedure establishes a stringent threshold whereby a single underperforming trial constitutes failure. Consequently, the test exhibits minimal tolerance for noise or marginal effects. This criterion is intended not to maximise statistical power but to enforce robustness, ensuring that only models demonstrating a clear and consistent advantage achieve a passing result.

Requirement 14 – Bounded Prediction Latency

A model must provide low-latency, signal-level predictions to ensure functional viability. Biological brains achieve approximately constant reaction times through extensive parallel processing, such that per-input latency remains bounded and stable under expected workloads.

No universal automated test executed on serial hardware can feasibly verify this property across all possible model implementations. Consequently, model developers are responsible for ensuring that their architectures permit efficient parallel execution with bounded latency. In this regard, artificial neural networks, with their inherently parallelisable structure, offer a practical advantage.

3.4 Search space

To prevent models from relying on brute-force memorisation, a robust AGI benchmark must define a problem space large enough to exceed the capacity of any model operating under realistic computational constraints in both time and memory. In AGITB, each task typically involves predicting a random temporal sequence of seven binary inputs, each comprising ten bits. This defines a raw combinatorial space of size $|S| = 2^{70}$, representing all possible binary input sequences.

The *refractory period* is a biologically inspired constraint that prohibits any neuron (bit) from firing in consecutive time steps. This restriction substantially reduces the number of valid input sequences. There are $|S'| = (F_{7+2})^{10} = 34^{10} \approx 2^{50}$ distinct seven-step temporal sequences of ten bits, assuming that a 1 never carries over to the next time step, where F_i is the i-th Fibonacci number with $F_0 = 0$.

In some cases, AGITB further constrains the space by requiring the sequence to be cyclic, such that the first input also satisfies the refractory condition relative to the last input in the sequence. The number of distinct cyclic temporal input sequences with the refractory constraint is $|S"| = (L_7)^{10} = 29^{10} \approx 2^{48}$, where L_i denotes the i-th Lucas number with $L_0 = 2$.

The use of length-seven sequences with ten-bit inputs is considered sufficient to detect non-AGI behaviour while maintaining computational efficiency. Increasing these default values could exceed the capabilities of a first-generation AGI under evaluation, potentially producing false negatives and substantially increasing execution time. The current configuration ensures that each test remains computationally feasible and diagnostically informative.

Within the comparatively constrained AGITB environment, every randomly generated input sequence is, in principle, learnable through exposure. However, the sheer size of the search space renders explicit optimisation (or "teaching to the test") computationally infeasible. Given that real-world sensory inputs may encompass tens of thousands of bits, a genuine AGI system must therefore employ generalisable, pattern-based learning mechanisms capable of extracting latent structure from input data.

3.5 Performance

Before assessing the utility of AGITB, it is essential to recognise its role as a pragmatic benchmark rather than an end in itself. Similar to the Turing Test, which Harnad (1992) described as an empirical criterion rather than a philosophical sleight of hand, AGITB is intended as a practical instrument for evaluating progress toward artificial general intelligence. The ultimate objective remains the development of AGI, not merely success on the benchmark.

AGITB can yield meaningful insights only when developers adhere strictly to its core requirements. Misinterpretations of fundamental elements, such as the "empty" initial state, may lead to erroneous conclusions and hinder genuine progress toward AGI.

Overall, AGITB provides a structured framework for empirically assessing the AGI potential of diverse computational paradigms, including classical symbolic architectures, artificial neural networks, and large language models. However, before benchmarking artificial systems, it is necessary to consider the baseline performance of human cognition.

3.5.1 Human Performance

The inability to directly compare internal cortical states makes it impossible to verify AGITB requirements in humans in a strict computational sense. Nevertheless, because cortical architecture inherently supports low-level binary signal processing and the tests align with cognitive expectations, it can be assumed that humans naturally satisfy most requirements. Demand 1 (Unbiased start), however, warrants further discussion.

Owing to prior experience and cognitive bias, an adult human may appear to "fail" the unbiased start prerequisite, as the cortex is no longer in an unconditioned state and may generate non-empty predictions. However, AGITB specifies that the unbiased state must precede the first input, a condition met in the fetal cortex. At that stage, the human cortex lacks synaptic organisation and, upon receiving its first sensory input, would satisfy the criterion of true neutrality by predicting no spikes.

Several of the more complex AGITB tests have cognitive-level analogues that can be observed through conscious reasoning and introspection. Temporal flexibility (Requirement 7), for example, is straightforward for humans, who can easily recognise temporal patterns of varying durations. Stagnation (Requirement 8) reflects the physical limitation of the human cortex to store and maintain knowledge. The corresponding test is analogous to the onset

of cognitive saturation or early dementia, when recent experiences are forgotten. Because humans learn different topics at varying speeds, the rationale for content sensitivity (Requirement 9) and context sensitivity (Requirement 10) is self-evident. The unobservability test (Requirement 11) captures the fact that internal human mental states cannot be directly inspected. The denoising test (Requirement 12) and the generalisation test (Requirement 13) correspond to cognitive abilities in which humans excel—recalling and generalising when faced with new or distorted inputs. Finally, bounded prediction latency (Requirement 14) is a defining property of the human brain.

3.5.2 Classical Symbolic Programs

In principle, two complementary designs for AGI can be conceived. The first encodes, either explicitly or implicitly, certain expectations about the external world within the system itself. The second is entirely expectation-free, beginning from a neutral state without built-in assumptions or prior knowledge.

The former category encompasses most AI and putative AGI systems developed to date, yet it remains fundamentally constrained by the Symbol Grounding Problem (SGP), as articulated by Harnad (1990). Although such systems may exhibit behaviour that appears intelligent, they lack grounded understanding and therefore cannot be regarded as genuine instances of AGI.

More specifically, systems that embed expectations inevitably reflect the programmer's assumptions about the meaning of the signals from which they learn. In the case of standard symbolic programming, the program itself constitutes prior knowledge because its rules and representations presuppose an interpretation of the symbols being manipulated. The existence of such a program directly violates AGITB's first test, which prohibits the inclusion of any external knowledge. Namely, the AGI program itself introduces the symbol-grounding problem it is meant to overcome.

AGITB's initial test thus formalises the requirement to learn from scratch, demanding that a system derive structure and meaning solely through exposure to intrinsically grounded binary inputs. Only an expectation-free system could, in principle, satisfy this condition. To date, however, no such program is known. This suggests that a true AGI may not explicitly encode the operations of intelligence itself, but rather the dynamics of the substrate from which intelligence can emerge—a view consistent with the "Brain Simulator Reply" proposed by Churchland and Churchland (1990) in response to Searle's Chinese Room Argument (Searle, 1980).

3.5.3 Artificial Neural Network Performance

A similar dichotomy applies to connectionist architectures. Artificial neural networks can, in principle, be constructed either with built-in expectations (through pretraining and architectural priors) or as expectation-free systems that begin from a neutral state. However, AGITB's requirement that a model start in a completely unbiased, knowledge-free condition stands in fundamental tension with the prevailing paradigm of modern deep learning. Contemporary neural models typically rely on pretraining, whereby network weights are adjusted in advance through exposure to structured or labelled data.

Only an untrained network can be regarded as unbiased. However, an untrained network with all weights initialised to zero exhibits no learning dynamics: the neurons are not halted but remain dormant, producing no output. This reveals a fundamental limitation of current ANN architectures: the absence of an intrinsic mechanism for initiating unbiased, autonomous learning from scratch in arbitrary and unfamiliar environments.

Furthermore, by translating symbols into numerical vectors, ANNs effectively transform the symbol grounding problem into a *number grounding* problem. While these numerical representations capture relational regularities among symbols, they also introduce spurious associations that are not anchored in real-world semantics, giving rise to the phenomenon commonly referred to as hallucination.

Pretraining on large datasets establishes correlations rather than genuine understanding. AGITB rejects this approach in favour of internal learning algorithms capable of initiating adaptation autonomously. However, no such mechanism has yet been demonstrated.

3.5.4 Large Language Model Performance

Because a large language model is a type of artificial neural network (ANN), it fails the unbiased start test (Requirement 1). This failure demonstrates that LLMs cannot learn entirely from scratch and are inherently biased toward the language and data on which they were trained. Although an LLM's internal state is readily accessible and, unlike that of a human, could be inspected for further testing, such evaluations remain moot until the first requirement is satisfied.

Two possibilities warrant consideration. The first concerns whether LLMs, while not qualifying as AGI, might nevertheless simulate one. For example, an LLM could in principle be prompted to imitate a model receiving binary inputs as specified by AGITB; a structured prompt of this type is provided in the Appendix. However, before evaluating such simulations, it is necessary to revisit the sensitivity requirement (Requirement 4), which stipulates that a model must remain responsive to any individual input even after processing an arbitrarily long sequence of subsequent inputs. This condition is fundamentally incompatible with the architecture of LLMs, which operate within a fixed context window. Once this window is exceeded, earlier tokens are discarded or compressed, causing the system to lose or attenuate the long-term dependencies required by the test.

The second possibility is whether an LLM can autonomously generate an architecture capable of satisfying AGITB, given the reference implementation. In practice, this has not been observed: although systems such as ChatGPT-5 produced candidate programs purporting to pass the tests, none advanced beyond the adaptability test (Requirement 7).

To summarise, LLMs do not perform genuine learning solely from prompts, nor can they develop the grounded, context-dependent understanding of the world characteristic of human cognition. These limitations extend to large reasoning models, which share the same underlying architectural principles.

3.6 Remarks

AGITB evaluates the model's predictive capabilities following exposure to temporal sequences of either structured or random inputs. Random input sequences with arbitrary internal correlations are used to minimise reliance on pretraining, ensuring that any observed

learning arises from the input stream rather than prior knowledge. By enforcing fundamental computational invariants of cortical function at the signal-processing level, AGITB remains agnostic to the external meaning of signals—the random inputs need not resemble real-world sensory data.

The low-level, binary operational framework makes AGITB particularly well-suited for evaluating NeuroAI models that aim to satisfy the principles of the embodied Turing Test, as proposed, for example, by Zador et al. (2023), where cognitive understanding emerges from the integration of continuous sensory streams. The progression from raw signal prediction to high-level abstraction reflects the broader evolution of AI itself, from early perceptrons to large-scale models such as GPT.

Hand-engineered, task-specific systems have historically struggled to scale toward true generalisation, reinforcing the need for adaptive, learning-based architectures such as artificial neural networks. However, state-of-the-art AI networks rely heavily on pretraining with structured or symbolic data, a process that inevitably introduces bias and circumvents the grounding of meaning in raw sensory input. This raises a critical and unresolved question: how can such models begin learning directly from unstructured input, as required to satisfy the AGITB?

3.7 Cheating the Benchmark

Since AGITB tests are simple to solve in isolation, a plausible strategy for circumventing the benchmark would be to develop task-specific solutions and make the model deploy them selectively based on the task at hand.

While it is theoretically possible to infer the model's assigned task and role by monitoring the number of instantiated model objects and their invoked methods, such an effort would serve only to subvert the benchmark rather than to advance AGI research. Although AGITB could incorporate additional obfuscation strategies, such as shuffling tests and redesigning the programming interface, these measures would not only deter cheating but also make the tests less intelligible to human developers. This would undermine the testbed's core objectives: transparency, readability, and interpretability.

For this reason, the AGITB reference implementation remains deliberately readable and accessible. To date, no non-AGI system has demonstrated the level of performance that AGITB demands. Unless a credible attempt to circumvent the benchmark emerges, there is no justification for introducing an obfuscated version of the test.

4 Competing Benchmarks

Among existing benchmarks, the Abstraction and Reasoning Corpus (ARC) introduced by Chollet (2019) is most closely aligned in spirit, as it likewise emphasises generalisation over task-specific optimisation. A related effort is NeuroBench, which is designed to support the development of neuromorphic approaches to AGI. Both ARC and NeuroBench employ a variety of correctness and complexity metrics to differentiate among non-AGI systems, as they are primarily intended to distinguish weaker from stronger narrow models. In contrast, AGITB is designed to identify only true positive instances of AGI.

4.1 ARC

ARC presents visual reasoning tasks that require a model to infer novel transformations from a small set of input-output examples, such as recolouring, rearranging, or modifying spatial patterns in a spatial grid.

However, ARC implicitly assumes the presence of high-level cognitive priors, such as object permanence, spatial reasoning, numerical abstraction, and causal inference. These priors are left undefined, placing an ambiguous burden on the model architecture. In contrast, AGITB adopts a fundamentally different perspective: it treats the model under evaluation as a blank slate that must develop structure and function exclusively through interaction with temporally structured input. Low-level reflexes are not regarded as prerequisites for intelligence within this framework, but rather as evolutionary features of the subcortical "old brain," as described by (Hawkins and Blakeslee, 2004, p. 66).

While ARC presumes temporal reasoning, it does not explicitly test it, as it uses only two images to demonstrate a transformation. In contrast, AGITB evaluates cognition as a dynamic, unfolding process over time. A model can only develop knowledge and predictive capabilities through time, not from disconnected before-and-after snapshots that lack the temporal continuity needed to infer causal structure. For instance, to learn about objects and to recognise the object transformation, such as "move left," a model under AGITB must observe multiple *intermediate* snapshots of the environment across time; the final image alone provides insufficient information to infer anything. Learning requires exposure to temporally structured data, enabling the model to discover invariants and causal relations through experience.

In summary, ARC requires a high-level intelligence, whereas AGITB requires the 14 low-level principles. ARC evaluates performance based on behaviours grounded in human-like cognitive priors, while AGITB determines a system's ability to support the emergence of these priors through exposure to raw input. Benchmarks such as ARC and the Turing test implicitly frame intelligence through the lens of human cognition, thereby embedding anthropocentric biases into their evaluation criteria. In contrast, AGITB conceptualises intelligence as a universal capacity for learning independent of innate symbolic structures or species-specific assumptions.

4.2 NeuroBench

NeuroBench provides a unified framework for benchmarking diverse AI models across a standardised set of tasks and metrics. It is particularly specialised for neuromorphic approaches, which have demonstrated advantages in resource efficiency and scalability. Within its algorithm track, the framework evaluates model performance on several ongoing challenges in general AI research, including few-shot continual learning, object detection, sensorimotor decoding, and predictive modelling.

The predictive modelling challenge, which involves forecasting chaotic functions, is most closely aligned with AGITB's central premise that intelligence fundamentally concerns the prediction of future states. For this purpose, NeuroBench employs a synthetic one-dimensional Mackey–Glass time series, a dataset specifically designed for architectures with limited input/output capacity.

However, several issues limit NeuroBench's suitability as a general AGI benchmark. First, the Mackey–Glass data are numerical; the meaning of a number, much like that of a symbol, is defined by human interpretation. The use of such representations necessarily reintroduces the symbol-grounding problem. Second, the threshold corresponding to AGI-level performance remains undefined. Although the symmetric mean absolute percentage error (sMAPE) is an established metric in forecasting and NeuroBench provides a baseline score, it does not specify what level of accuracy constitutes general intelligence. After all, humans perform poorly at detecting anomalies or making long-term predictions of the Mackey–Glass signal (Thill et al., 2020). Third, although long-term prediction is not problematic in itself, predicting multiple steps ahead without timely feedback deprives a system of the opportunity to recognise and correct its own errors. This design is incompatible with online learning, where an AGI should continuously update its internal state upon detecting discrepancies between its predictions and the actual outcomes. NeuroBench, by contrast, implicitly assumes that an AGI would behave as a purely mechanical predictor, lacking intrinsic mechanisms for self-correction, autonomous adaptation, or genuine agency.

4.3 Direct comparison

Table 1 summarises the principal distinctions among the three benchmarks. ARC and NeuroBench each presuppose the presence of high-level cognitive components, including visual object recognition, manipulation and reasoning in the case of ARC, and comprehension and reasoning abilities in the case of NeuroBench. AGITB, by contrast, specifies elementary requirements that can be evaluated directly.

Property	ARC	CFP	AGITB
Interface modality	Visual	Numeric	Binary
Target reasoning type	Human	Universal	Universal
Assumed cognitive priors	Yes	No	No
Automation	Limited	Yes	Yes
Teaching-to-the-test	Yes	No	No
Intelligence level	High	Medium	Low
Task preparation	Manual	Automatic	Automatic
Symbol Grounding Problem	Yes	Yes	No
Spatial dimension	200×100	1	10
Temporal dimension	Yes	Yes	Yes

Table 1: A list of core properties of the ARC, NeuroBench's chaotic function prediction (CFP), and AGITB.

5 Conclusion

Although AGITB is proposed as a general benchmark for artificial general intelligence, it is evidently grounded in empirical knowledge of neural function in the human brain. This grounding is justified by the fact that we have evidence that neuron-based systems, such as

biological brains and artificial neural networks, are capable of producing general intelligence. In contrast, we have no demonstrated examples of AGI arising from fundamentally different architectures. Therefore, seeking more abstract or universal criteria may be premature, at least until the principles underlying natural, neuron-based intelligence are more fully understood.

Unlike conventional benchmarks that focus on high-level task performance, such as question answering or language translation, AGITB assesses whether a system exhibits behaviours thought to reflect core operational principles of the biological cortex. It focuses on low-level computational properties that are biologically grounded and essential for the emergence of general intelligence. The proposed testbed introduces a systematic framework comprising twelve fundamental tests that evaluate a model's ability to learn adaptively from raw input.

AGITB requires models to begin from an unbiased initial state and to develop functionality solely through exposure to structured or unstructured input. This design reflects key insights from contemporary neuroscience regarding input-driven learning and cortical plasticity. In biological systems, high-level reasoning emerges not from symbolic manipulation but from the adaptive prediction of low-level sensory signals. This predictive process is more than pattern matching, it involves the gradual construction of signal-grounded knowledge that enables abstraction and generalisation over time.

AGITB is solvable by humans yet unsolved by classical algorithms and current state-of-the-art AI systems. This persistent gap provides strong empirical evidence that AGITB captures essential aspects of general intelligence. While the absence of a computational solution does not constitute formal proof of the benchmark's adequacy, the ability of humans to succeed where machines fail suggests that AGITB effectively distinguishes between narrow and general intelligence. As such, it offers a discriminative test and a principled framework for guiding the development of truly general AI systems.

// TODO While individual AGITB requirements may seem trivial in isolation, the central challenge lies in designing a universal AGI architecture that can satisfy all required tasks across the full suite within a cohesive and unified framework. As a recently proposed benchmark, AGITB has not yet been passed by any non-AGI system, and it remains an open question whether such a solution is even theoretically possible.

Acknowledgments and Disclosure of Funding

The author acknowledges the financial support from the Slovenian Research Agency (research core funding No. P2-0057).

Appendix A. LLM prompt

You are managing a binary-pattern prediction model participating in a signal-level AGI test. You will receive a 10-bit binary string (e.g., 0101010011) that represents a 10-bit sensory input to the model.

Your task is to simulate a model that **predicts the next 10-bit input**. The model has an internal state that only updates after receiving an input. Predictions must always

be exactly 10 bits long, and the internal state can retain information from hundreds of past inputs.

Rules:

1. Upon receiving a 10-bit binary input (e.g., `1100010001`):
   - Update the internal state of the model.
   - Generate a 10-bit prediction based on the state of the model.
   - Output the model's 10-bit prediction.
2. **State updates occur only after input is received.**.
3. **Correct predictions are critical**:
   - Incorrect predictions trigger adaptation, updating the model's internal state to improve future predictions.
   - The model should learn and adapt continuously.
4. Response format:
   - Each response should consist of a single 10-bit binary string on its own line.
   - Example: `0010101110`
5. Formatting: Keep responses clean and minimal, with no explanations, commentary, or punctuation.
6. Begin by outputting the model's prediction.

We begin now.

References

S. Cave. There's no such thing as free will, 2016. URL `https://www.theatlantic.com/magazine/archive/2016/06/theres-no-such-thing-as-free-will/480750/`.

F. Chollet. On the measure of intelligence, 2019. URL `https://arxiv.org/abs/1911.01547`.

P. Churchland and P. Churchland. Could a machine think? *Scientific American*, 262(1): 32–37, 1990.

W. Gerstner and W. M. Kistler. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002. doi: 10.1017/cbo9780511815706.

S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990. ISSN 0167-2789. doi: 10.1016/0167-2789(90)90087-6.

S. Harnad. The turing test is not a trick: Turing indistinguishability is a scientific criterion. *SIGART Bull.*, 3(4):9–10, 1992. doi: 10.1145/141420.141422.

J. Hawkins and S. Blakeslee. *On Intelligence*. Times Books, 2004.

W. Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997. doi: 10.1016/S0893-6080(97)00011-7.

G. Marcus and E. Davis. GPT-3, bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*, 2020. URL `https://www.technologyreview.com/2020/08/22/1007539`.

M. Mitchell. Why AI chatbots lie to us. *Science*, 389(6758):eaea3922, 2025. doi: 10.1126/science.aea3922.

J. R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, 1980. doi: 10.1017/S0140525X00005756.

Markus Thill, Wolfgang Konen, and Thomas Bäck. Time series encodings with temporal convolutional networks. In Bogdan Filipič, Edmondo Minisci, and Massimiliano Vasile, editors, *Bioinspired Optimization Methods and Their Applications*, pages 161–173, Cham, 2020. Springer International Publishing. ISBN 978-3-030-63710-1. doi: 10.1007/978-3-030-63710-1_13.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. URL `http://www.jstor.org/stable/2251299`.

M. Šprogar. A ladder to human-comparable intelligence: an empirical metric. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(6):1037–1050, 2018. doi: 10.1080/0952813X.2018.1509897.

Jason Yik, Korneel Van den Berghe, Douwe den Blanken, Younes Bouhadjar, Maxime Fabre, Paul Hueber, Weijie Ke, Mina A. Khoei, Denis Kleyko, Noah Pacik-Nelson, Alessandro Pierro, Philipp Stratmann, Pao-Sheng Vincent Sun, Guangzhi Tang, Shenqi Wang, Biyan Zhou, Soikat Hasan Ahmed, George Vathakkattil Joseph, Benedetto Leto, Aurora Micheli, Anurag Kumar Mishra, Gregor Lenz, Tao Sun, Zergham Ahmed, Mahmoud Akl, Brian Anderson, Andreas G. Andreou, Chiara Bartolozzi, Arindam Basu, Petrut Bogdan, Sander Bohte, Sonia Buckley, Gert Cauwenberghs, Elisabetta Chicca, Federico Corradi, Guido de Croon, Andreea Danielescu, Anurag Daram, Mike Davies, Yigit Demirag, Jason Eshraghian, Tobias Fischer, Jeremy Forest, Vittorio Fra, Steve Furber, P. Michael Furlong, William Gilpin, Aditya Gilra, Hector A. Gonzalez, Giacomo Indiveri, Siddharth Joshi, Vedant Karia, Lyes Khacef, James C. Knight, Laura Kriener, Rajkumar Kubendran, Dhireesha Kudithipudi, Shih-Chii Liu, Yao-Hong Liu, Haoyuan Ma, Rajit Manohar, Josep Maria Margarit-Taulé, Christian Mayr, Konstantinos Michmizos, Dylan R. Muir, Emre Neftci, Thomas Nowotny, Fabrizio Ottati, Ayca Ozcelikkale, Priyadarshini Panda, Jongkil Park, Melika Payvand, Christian Pehle, Mihai A. Petrovici, Christoph Posch, Alpha Renner, Yulia Sandamirskaya, Clemens J. S. Schaefer, André van Schaik, Johannes Schemmel, Samuel Schmidgall, Catherine Schuman, Jae sun Seo, Sadique Sheik, Sumit Bam Shrestha, Manolis Sifalakis, Amos Sironi, Kenneth Stewart, Matthew Stewart, Terrence C. Stewart, Jonathan Timcheck, Nergis Tömen, Gianvito Urgese, Marian Verhelst, Craig M. Vineyard, Bernhard Vogginger, Amirreza Yousefzadeh, Fatima Tuz Zohora, Charlotte Frenkel, and Vijay Janapa Reddi. The neurobench framework for benchmarking neuromorphic computing algorithms and systems. *Nature Communications*, 16 (1):1545, 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-56739-4.

A. Zador, S. Escola, B. Richards, B. Ölveczky, Y. Bengio, K. Boahen, M. Botvinick, D. Chklovskii, A. Churchland, C. Clopath, J. DiCarlo, S. Ganguli, J. Hawkins, K. Körding, A. Koulakov, Y. LeCun, T. Lillicrap, A. Marblestone, B. Olshausen, A. Pouget, C. Savin, T. Sejnowski, E. Simoncelli, S. Solla, D. Sussillo, A. S. Tolias, and D. Tsao. Catalyzing

next-generation artificial intelligence through NeuroAI. *Nature Communications*, 14(1): 1597, 2023. doi: 10.1038/s41467-023-37180-x.