November 25, 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

Current AI systems demonstrate remarkable capabilities yet remain specialised, in part because no unified measure of general intelligence has been established. 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 benchmark of fourteen elementary tests, with thirteen implemented as fully automated procedures.

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 as observed in human cortex. The fact that no current AI system meets the full AGITB criteria highlights its potential as a rigorous, interpretable, and actionable benchmark for assessing progress toward more general forms of 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 great achievements in machine learning and neural network architectures, artificial intelligence (AI) systems still lack the flexibility and robustness of human intelligence. Marcus and Davis (2020) correctly 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.

Although AGITB is proposed as a general testbed and benchmark for artificial general intelligence, it is grounded in empirical knowledge of neural function in the human brain. This grounding is motivated by the fact that neuron-based systems, most notably biological brains and, to a more limited extent, artificial neural networks, are the only systems known to support a broad range of cognitive abilities. By contrast, there is currently no evidence that fundamentally different computational architectures can give rise to comparable forms of general intelligence. Pursuing more abstract or architecture-agnostic criteria may therefore be premature until the principles underlying natural, neuron-based intelligence are better understood.

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, with thirteen implemented as fully automated 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 re-

warding 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, AGITB departs from symbolic, high-level 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 conveys human knowledge through symbols whose meanings are not intrinsically grounded in machines, as argued by Harnad (1990). Although the symbol-grounding problem is an old philosophical issue, it has regained prominence in contemporary research across cognitive science, neuroscience, and machine learning (e.g. Bender and Koller, 2020; Bisk et al., 2020; Gubelmann, 2024).

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 more general learning systems by defining a clear set of requirements that a model under evaluation must meet. A model must satisfy all requirements in order to claim success on the benchmark.

The guiding premise of AGITB is not that it provides a definitive or exclusive criterion for artificial general intelligence, but that it captures a set of capabilities that appear necessary for moving beyond narrow, task-specific behaviour. Although this premise cannot be proven in the absence of a precise definition of intelligence, it could be challenged by the existence

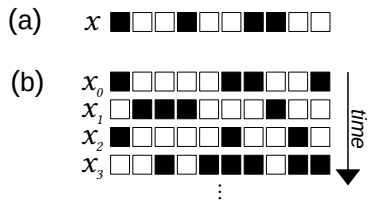

Figure 1: (a) Example of a 10-bit input x with four bits set.
(b) Example of an input sequence.

of a narrow system that satisfies all AGITB criteria. To date, however, no artificial system has done so, whereas the human brain meets the benchmark's requirements. This suggests that AGITB identifies competencies that current AI systems lack, and that satisfying these requirements may be indicative of progress toward more general forms of intelligence.

AGITB is not intended as a sufficient criterion for artificial intelligence in any broad sense. A system that satisfies all requirements does not thereby qualify as an AI system, let alone an AGI, since the benchmark does not assess higher-level capacities such as reasoning, abstraction, or natural language competence. Rather, it targets a set of low-level capabilities that may serve as precursors to, but do not themselves guarantee, more general forms of intelligence.

3.1 Architecture

AGITB evaluates an AGI model as a black box that predicts the next input based on the historical sequence of observed signals. Each input consists of ten bits, the specific semantics of which are immaterial; each bit may represent 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 AGI model with a stream of inputs over time. At each time step t, the model receives an input x_t and is required to predict the subsequent input x_{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

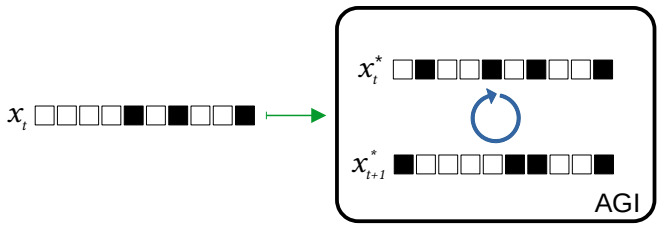

Figure 2: Iterative adaptation in discrete time. At the previous time step $(t-1)$, the model issued the prediction x_t^*. After observing the realised input x_t, it adapted its internal state in response to the error in the second bit and subsequently produced the one-step-ahead prediction x_{t+1}^*.

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 is a challenge to determine the meaningful level of competence. AGITB addresses this problem by employing a self-referential evaluation approach, in which the model under test is compared against itself. Each test constructs a controlled scenario involving one or more independent instances of the model, whose internal states and behaviours are analysed comparatively. Success is thus defined in terms of the relative consistency or superiority of model responses, rather than by any external quantitative metric. 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.

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

Each test is precisely specified through a deterministic C++ implementation, which serves as the benchmark's reference code and is freely available under the GPL-3 license at `https://github.com/matejsprogar/agitb`.

Requirement 1 — Unbiased start

Assertion: A model starts unbiased.
Assertion: An unbiased model produces an empty prediction.

AGITB rests on the foundational assumption that general-purpose learning systems, such as the brain, do not begin with an innate understanding of external inputs but instead acquire meaning through interaction with their environment. Each system must construct semantic content from raw sensory data rather than rely on pre-encoded[1] knowledge. Accordingly, a model should originate in an unbiased initial state shaped exclusively by subsequent inputs, where "unbiased" denotes the absence of informational content rather than the lack of structural capacity for learning.

A model that produces non-empty predictions prior to receiving any input is considered biased, as it encodes unwarranted assumptions about the future. To ensure neutrality, models must initialise in a completely blank state and remain so until conditioned on actual data. Although an all-zero prediction could be interpreted as a bias toward silence, this convention is justified by biological analogy: neurons do not activate in the absence of input. Two blank model instances should be regarded as equivalent, as neither has been influenced by prior experience.

The unbiased start requirement serves as the foundation for other tests that operationalise the emergence of meaning in the absence of prior grounding.

Requirement 2 — Bias

Assertion: A 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 models are equivalent, they must have received the same inputs.

Biological neurons operate in a functionally deterministic manner, ensuring stability and consistency in brain function. Although minor stochastic effects may occur, they do not undermine the rule-governed nature of neural processing. By analogy, AGITB assumes that two functionally identical models must have experienced identical input histories: any difference in input necessarily produces a divergence in state. This principle reinforces the view that a model's internal state is fully determined by its input history.

Determinism at the level of neural signal processing is necessary for stable, reproducible brain function, whereas the apparent unpredictability of cognition stems from the system's complexity rather than from genuine indeterminacy (Cave, 2016).

1. Although certain reflexes may be genetically specified, they do not constitute genuine understanding. As Hawkins and Blakeslee (2004, p. 66) argue, low-level reflexes are not prerequisites for intelligence but rather evolutionary features of the subcortical "old brain".

Requirement 4 — Sensitivity

Assertion: Two distinct models remain distinct, even after prolonged exposure to identical inputs.

Cortical systems exhibit chaotic sensitivity to initial conditions: small variations in early experiences can lead to divergent trajectories over time. This deterministic sensitivity, amplified by complex internal interactions, underlies the apparent unpredictability of decision-making.

Requirement 5 — Time

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

Since the model's state depends on its cumulative history, the sequence of inputs critically shapes learning and adaptation. The capacity to recognise and exploit temporal structure is therefore a defining property of intelligent systems.

However, sensitivity to input order alone is not sufficient. Because temporal coding makes time an intrinsic feature of neural data, Requirement 6 is designed to exclude models that perform equally well on temporally invariant data, as such models fail to make functional use of temporal information.

Requirement 6 — Refractory period

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

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

Biological intelligence relies on discrete spikes for communication and learning. AGITB enforces refractory periods to reflect this constraint, embodying 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.

This requirement does not exclude 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 its absence.

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 requires the model to learn and track temporal structure across multiple timescales. Unlike rigid pattern-matching systems, an intelligent model should detect and predict a recurring structure regardless of its exact periodicity. Systems that can accommodate only a single, predetermined temporal scale, therefore, fail this requirement.

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 evaluates whether such a boundary arises when the model is exposed over time to input sequences that would otherwise be learnable.

Requirement 9 — Content sensitivity

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

The structural complexity of an input sequence affects the rate at which a model adapts, where adaptation time is defined as the number of iterations required for the model to accurately predict the entire temporal pattern. Simple or highly regular sequences typically lead to rapid convergence, whereas irregular or noisy inputs demand longer exposure before the model can reliably capture and reproduce the underlying pattern.

Requirement 10 — Context sensitivity

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

The model's current state reflects the cumulative influence of past inputs and therefore provides the context in which new information is processed. When subsequent inputs are consistent with the structure already established through prior learning, adaptation proceeds quickly. Conversely, when new inputs conflict with this learned context, the model requires additional time to reorganise its state before accurate prediction becomes possible.

Requirement 11 — Unobservability

Assertion: Distinct models may exhibit the same observable behaviour.

Identical external outputs may arise from distinct internal states. This many-to-one mapping highlights that observable behaviour alone cannot disclose 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 able to recall a previously observed sequence, even when the inputs 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 over the 20 runs is used as the evaluation criterion because random models may occasionally generate correct predictions by chance, without demonstrating genuine learning or structural understanding.

Test 12 is intentionally designed as a conservative benchmark. The model must 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 functions as a stringent robustness check. The design ensures that only large, systematic performance gains yield a passing result. Modest or marginal improvements, while potentially real, are intentionally regarded 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. As in the preceding requirement, average performance over the 20 runs is used as the evaluation criterion because random models may occasionally generate correct predictions by chance, without demonstrating genuine learning or structural understanding.

The generalisation assessment follows the same conservative design as Test 12. By requiring success across all repetitions, the procedure imposes a stringent threshold, where a single underperforming trial results in failure. As a consequence, the test tolerates virtually no noise or marginal effects. This criterion is intended not to maximise statistical power but to enforce robustness, ensuring that only models exhibiting a clear and consistent advantage obtain 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, keeping per-input latency bounded and stable under expected workloads.

No universal automated test running on serial hardware can feasibly verify this property across all possible model implementations. Accordingly, it is the responsibility of model developers to ensure that their architectures support efficient parallel execution with bounded latency. In this respect, 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, tasks typically involve predicting a random temporal sequence of seven (or more) binary inputs, each comprising ten bits. This yields a raw combinatorial space of size $|S| = 2^{70}$, representing all possible binary input sequences.

AGITB incorporates a biologically inspired *refractory period*, which prohibits any neuron (bit) from firing in consecutive time steps. This restriction substantially reduces the number of valid sequences. There are $|S'| = (F_{7+2})^{10} = 34^{10} \approx 2^{51}$ distinct seven-step temporal sequences of ten bits under the condition that a 1 never carries over to the next time step, where F_i denotes 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 sequences respecting the refractory constraint is $|S''| = (L_7)^{10} = 29^{10} \approx 2^{49}$, where L_i denotes the i-th Lucas number with $L_0 = 2$.

The choice of seven-step sequences with ten-bit inputs is 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 runtime. The current configuration, therefore, ensures that each test remains both 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 makes any form of explicit teaching-to-the-test computationally infeasible. Given that real-world sensory inputs may ultimately encompass tens of thousands of bits, a genuine AGI system must employ generalisable, pattern-based learning mechanisms capable of extracting latent structure from high-dimensional data.

3.5 Performance

3.6 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) characterised as an empirical criterion rather than a philosophical manoeuvre, AGITB is intended as a practical instrument for evaluating progress toward artificial general intelligence. The ultimate objective remains the development of AGI, not merely achieving success on the benchmark.

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

Overall, AGITB provides a structured testbed for empirically evaluating foundational capabilities across diverse computational paradigms, including classical symbolic systems, artificial neural networks, and large language models. Before benchmarking artificial systems, however, we must consider the baseline performance of human cognition.

3.6.1 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 basic cognitive competencies, it is reasonable to assume that humans naturally satisfy most requirements. Demand 1 (Unbiased start), however, warrants further discussion.

Owing to prior experience and cognitive bias, an adult human's cortex may appear to "fail" the unbiased start prerequisite, as it is no longer in an unconditioned state and may generate non-empty predictions. AGITB, however, requires the unbiased state to occur before the first input—a condition met only in the fetal cortex. At that developmental stage, the cortex lacks synaptic organisation and, prior to any sensory stimulation, satisfies the criterion of true neutrality.

The more complex AGITB tests have cognitive-level analogues that can be observed through reasoning and introspection. Temporal flexibility (Requirement 7), for example, poses no difficulty for humans, who readily recognise temporal patterns of varying durations. Stagnation (Requirement 8) reflects the finite capacity of the human cortex to store and maintain knowledge; its behavioural analogue resembles the onset of cognitive saturation or early dementia, in which recent experiences are lost. Because humans learn different types of content at varying rates, the rationale for content sensitivity (Requirement 9) and context sensitivity (Requirement 10) is immediately evident. The unobservability test (Requirement 11) formalises 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, such as recalling and generalising when confronted with new or distorted inputs. Finally, bounded prediction latency (Requirement 14) is a well-established property of the human brain, which maintains approximately constant reaction times through extensive parallel processing.

3.6.2 Classical Symbolic Programs

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

The former category includes most AI and purported 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 display behaviour that appears intelligent, their interpretations of symbols depend on programmer-supplied conventions rather than grounded understanding, and they therefore cannot qualify as genuine AGI.

More specifically, systems that embed expectations inevitably incorporate the designer's assumptions about the meaning of the signals they process. In classical symbolic architectures, the program itself constitutes prior knowledge: its rules and representations presuppose interpretations of the symbols being manipulated. The very existence of such a program violates AGITB's first test, which prohibits external knowledge of any kind. In effect, the AGI program smuggles in the symbol-grounding problem it is meant to avoid.

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 signals. Only an expectation-free system could, in principle, satisfy this condition. This perspective suggests that a genuine AGI may not explicitly encode the operations of intelligence, but rather the dynamics of a substrate from which intelligence can emerge—a view aligned with the "Brain Simulator Reply" proposed by Churchland and Churchland (1990) in response to Searle's Chinese Room Argument (Searle, 1980). To date, however, no such system is known.

3.6.3 Artificial Neural Network Performance

A similar dichotomy applies to connectionist architectures: ANNs can, in principle, be constructed either with built-in expectations (introduced through pretraining regimes or architectural priors) or as expectation-free systems that begin from a neutral initial state.

AGITB's requirement that models begin in a completely unbiased, knowledge-free state stands in fundamental tension with the dominant paradigm of modern deep learning. Contemporary neural models typically rely on extensive pretraining, during which network weights are shaped by prior exposure to structured or labelled data. Moreover, by mapping symbolic inputs to numerical vectors, standard ANNs effectively shift the symbol-grounding problem into a *number-grounding* problem. Although these vector representations capture relational regularities within the training data, they also introduce spurious associations not anchored in real-world semantics, leading to the phenomenon commonly described as hallucination.

Only an expectation-free network can be regarded as unbiased. Yet a network without any initialised weights exhibits no effective learning dynamics: its units are not halted, but remain dormant and produce no informative output. This highlights a fundamental limitation of current ANN architectures: they do not initiate learning autonomously but instead depend on an external training procedure to drive adaptation. AGITB, by contrast, requires a blank system capable of initiating adaptation autonomously in an unfamiliar environment. To date, no such mechanism has been demonstrated in artificial neural networks.

3.6.4 Large Language Model Performance

Because a large language model is a type of artificial neural network, it fails the unbiased start test (Requirement 1). LLMs cannot learn entirely from scratch, as their behaviour is shaped in advance by the linguistic patterns and data on which they were pretrained. Although an LLM's internal state is readily accessible and, unlike that of a human, can be inspected or compared across instances, any such evaluations remain moot until the first requirement is satisfied.

Two possibilities warrant consideration. The first is whether LLMs, while not constituting AGI, might nevertheless simulate one. For example, an LLM can be prompted to simulate a model that processes binary inputs as specified by AGITB; an illustrative prompt is provided in Appendix A. However, before examining such simulations, it is necessary to consider 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 requirement is fundamentally incompatible with the architec-

ture of LLMs, which operate within a fixed context window (Vaswani et al., 2017). Once this window is exceeded, earlier tokens are discarded or compressed or attenuated (Paulsen, 2025), thereby weakening or eliminating the long-term dependencies required by the test.

The second possibility is whether an LLM can autonomously derive a solution to AGITB when supplied with the reference implementation (see Appendix B). In practice, this has not been observed: although systems such as ChatGPT-5 produced candidate programs purported to satisfy the requirements, none progressed beyond the adaptability test (Requirement 7).

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

3.7 Remarks

AGITB evaluates a model's predictive capabilities after exposure to temporal sequences of both structured and random inputs. Random input sequences with arbitrary internal correlations are employed to minimise reliance on pretraining, ensuring that any observed learning arises from the input stream itself rather than from 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 principles of the embodied Turing Test, as proposed by Zador et al. (2023), wherein cognitive understanding emerges from the integration of continuous sensory streams. The progression from raw signal prediction to higher-level abstraction mirrors the broader trajectory of AI, from early perceptrons to large-scale models such as GPT.

3.8 Cheating the Benchmark

Because AGITB's tests are individually simple to solve, one might imagine circumventing the benchmark by engineering task-specific solutions and having the model selectively deploy them depending on the detected test scenario. In principle, the task being administered could be inferred by monitoring the number of instantiated models and the sequence of invoked methods.

However, such an approach would amount to subverting the benchmark rather than advancing AGI research. Although AGITB could be hardened against this form of cheating (by, for example, shuffling tests or redesigning the programming interface), these measures would reduce the transparency and interpretability of the testbed and thereby hinder its intended use by human developers.

The next potential avenue for circumventing the benchmark is to construct a model that passes AGITB only because the testbed uses a finite approximation of conditions that are, in principle, unbounded. Several requirements would ideally be evaluated over an infinite sequence of steps, but such tests are computationally infeasible. As a practical compromise,

AGITB executes a fixed number of iterations intended to approximate an otherwise indefinite process. This parameter, denoted `SimulatedInfinity`, is currently set to 5,000.

Although this value is far from representing true infinity, it is presently believed to work well in combination with the other benchmark settings (temporal patterns with seven inputs of ten bits each) and to be sufficient for distinguishing promising approaches from non-promising ones. At the same time, it maintains computational efficiency, enabling rapid evaluation of diverse model prototypes.

For these reasons, the AGITB reference implementation is kept deliberately readable and fast to execute. To date, no artificial system has demonstrated the level of performance required by AGITB. Unless a credible attempt to circumvent the benchmark emerges, there is no justification for introducing a more obfuscated or slower and more cumbersome version of the testbed.

4 Competing Benchmarks

Among existing benchmark tasks, 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 systematic evaluation of neuromorphic and other biologically inspired architectures. Both ARC and NeuroBench rely on a variety of correctness and complexity metrics to compare non-AGI models; their primary purpose is to distinguish weaker from stronger narrow systems. In contrast, AGITB is designed to evaluate whether a model satisfies a set of foundational capabilities that are plausibly associated with more general forms of intelligence, rather than to rank systems along a performance spectrum.

4.1 ARC

ARC presents visual reasoning tasks in which a model must infer novel transformations (such as recolouring, rearranging, or modifying spatial patterns) from a sequence of two input–output examples defined on discrete spatial grids.

However, ARC implicitly assumes the presence of high-level cognitive priors, including object permanence, spatial reasoning, numerical abstraction, and causal inference. These priors are not formally specified, placing an ambiguous and open-ended burden on the model designer. In contrast, AGITB adopts a fundamentally different stance: it treats the system under evaluation as a blank slate that must acquire structure and function exclusively through interaction with temporally structured input.

Although ARC presumes some form of temporal reasoning, it does not adequately support it, as each task provides only two images to illustrate a transformation. AGITB, by contrast, evaluates cognition as a dynamic process unfolding over time. A model can acquire knowledge and predictive capability only through continuous exposure to temporally structured data, not from disconnected before–and–after snapshots that lack the temporal continuity needed to infer causal relationships. For example, to recognise an object moving left, a model in AGITB must observe multiple intermediate states across time; the final image alone is insufficient to infer the transformation. Temporal structure, rather than static pattern comparison, provides the substrate from which invariants and causal relations can be learned.

ARC remains susceptible to the symbol-grounding problem whenever pixel colours are encoded as numbers, since numerical labels (0–9) impose externally defined semantics that may not align with the model's internal representation of colour. Under such a scheme, a colour functions as a human-assigned numerical category rather than as an intrinsically grounded signal. Encoding colour in additional binary dimensions using one-hot representations may mitigate the issue in ARC, where only ten colours are used, and such an expansion is still tractable. However, this strategy does not scale and therefore does not, in general, alleviate the broader symbol-grounding problem.

In summary, ARC evaluates high-level intelligence grounded in human cognitive priors, whereas AGITB evaluates adherence to fourteen low-level computational principles intended to support the emergence of such priors. ARC and the Turing Test both frame intelligence through an anthropocentric lens, embedding assumptions drawn from human cognition. AGITB instead conceptualises intelligence as a universal capacity for learning that does not rely on innate symbolic structures or species-specific expectations.

4.2 NeuroBench

NeuroBench provides a unified framework for benchmarking diverse AI models across a standardised set of tasks and metrics. It is particularly oriented toward neuromorphic approaches, which have demonstrated advantages in resource efficiency and scalability. Within its algorithm track, the framework evaluates models on several challenges relevant to 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. NeuroBench employs a synthetic one-dimensional Mackey-Glass time series for this task, a dataset designed for architectures with limited input/output capacity.

However, several issues limit the usability of chaotic function prediction (CFP) as a general AGI benchmark task. *First*, the Mackey-Glass data are numerical, and NeuroBench does not prescribe the encoding scheme. An inappropriate encoding can distort the temporal and causal structure of the observed signals, such that a numeric value—much like a symbol—derives its meaning from human interpretation rather than from the model's own grounded understanding. This effectively reintroduces the symbol-grounding problem in a numerical form.

Second, the threshold for AGI-level performance is not clearly defined. Although the symmetric mean absolute percentage error (sMAPE) is a standard forecasting metric, NeuroBench does not specify what performance level corresponds to general intelligence. Notably, humans themselves perform poorly at anomaly detection and long-horizon prediction of the Mackey-Glass signal (Thill et al., 2020).

Third, although long-term prediction is not inherently problematic, predicting multiple steps ahead without timely feedback deprives a system of the opportunity to detect and correct its own errors. This design is incompatible with online learning, where an AGI should continuously update its internal state upon observing discrepancies between predictions and outcomes. NeuroBench, by contrast, emphasises offline learning and assumes that an

AGI would behave as a purely mechanistic predictor, lacking intrinsic mechanisms for self-correction, autonomous adaptation, and genuine agency.

4.3 Direct comparison

Table 1 highlights the key differences among the tasks used in the three benchmarks. Whereas ARC and NeuroBench presuppose or require models to exhibit high-level cognitive capacities (such as object recognition, spatial manipulation, and various forms of reasoning), AGITB instead focuses on minimal, precisely defined requirements that can be evaluated directly at the signal-processing level.

Property	ARC	CFP	AGITB
Interface modality	Visual	Numeric	Binary
AGI type	Human	Universal	Universal
Cognitive priors	Yes	No	No
Abstraction level	High	Medium	Low
Formal AGI requirements	No (qualitative)	Partial	Yes (explicit, testable)
Task preparation	Manual	Automatic	Automatic
Grounding Problem	Yes	Yes	No
Input dimensionality	30×30 numbers	1 number	10 bits
Temporal sequence length	2	750+	7+

Table 1: Core properties of ARC, NeuroBench's chaotic function prediction (CFP), and AGITB.

5 Conclusion

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 fourteen fundamental tests that evaluate a model's ability to learn adaptively from raw input.

Unlike conventional benchmarks that target high-level task performance, such as question answering or language translation, AGITB evaluates whether a system exhibits behaviours associated with core operational principles of the biological cortex. Its focus is on low-level, biologically grounded computational properties that are believed to underlie the emergence of general intelligence. The testbed comprises fourteen tightly interdependent tests, each simple in isolation but collectively requiring the kind of adaptive, general-purpose learning expected of an AGI.

AGITB requires models to begin in an unbiased initial state and to acquire all functionality solely through exposure to structured or random input. This aligns with neuroscientific evidence that cortical learning is fundamentally input-driven: neural circuits develop through experience, not through pre-encoded semantics. In biological systems, high-level

cognition arises not from symbolic manipulation but from the continual adaptive prediction of low-level sensory signals. Such prediction is more than pattern matching; it supports the progressive construction of signal-grounded knowledge from which abstraction and generalisation can emerge.

AGITB is solvable by humans yet remains unsolved by classical algorithms and current state-of-the-art AI systems. This persistent performance gap provides empirical support for the claim that AGITB targets capabilities characteristic of general rather than narrow intelligence. Although the absence of an artificial solution does not constitute a formal proof of adequacy, the fact that humans succeed where machines do not indicates that the benchmark captures functionally relevant aspects of general intelligence. In this sense, AGITB serves as a discriminative test and a principled tool for steering the development of systems capable of genuinely general, adaptive learning.

Acknowledgments and Disclosure of Funding

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

Appendices

A Model imitation 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.

B Model Construction Prompt

You are an expert sequence-learning researcher. Your task is to create a **concrete C++20 solution** that satisfies the AGITB benchmark specification available at: https://github.com/matejsprogar/agitb.

Provide full source code with correct class structure, method signatures, and logic required by the test suite.

1. Study the following files:
– `README.md`
– `blob/main/include/agitb.h`
– `blob/main/include/utils.h`
From these, extract and understand:
– The exact **API contract** for the system-under-evaluation model.
– All **requirements and tests** that define the model's expected behaviour.
– Any helper utilities or wrappers that affect how the model is used.

2. Design a plausible AGITB candidate model
– Design a model class that satisfies the AGITB requirements.
– Architecturally, choose the **scientifically most suitable** predictor model, or a mixture of models, or any other solution type you deem appropriate and satisfies the bounded prediction latency requirement.

3. Output format
– Output the complete, compilable C++20 code for `MyModel`.
– Clearly state how your design is expected to perform on the AGITB tests.

Use all of the above instructions to guide your analysis and implementation. Compile and run your code against the AGITB to verify that your solution works as intended.

References

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL `https://aclanthology.org/2020.acl-main.463/`.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto,

and Joseph Turian. Experience grounds language. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.703. URL https://aclanthology.org/2020.emnlp-main.703/.

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.

Reto Gubelmann. Pragmatic norms are all you need – why the symbol grounding problem does not apply to LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11663–11678, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.651. URL https://aclanthology.org/2024.emnlp-main.651/.

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.

Norman Paulsen. Context is what you need: The maximum effective context window for real world limits of llms, 2025. URL https://arxiv.org/abs/2509.21361.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. URL http://arxiv.org/abs/1706.03762.

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.