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

Despite major advances in machine learning, current artificial intelligence systems continue to fall short of human-like general intelligence. While large language and reasoning models can generate fluent and coherent outputs, they lack the deep understanding and adaptive reasoning that characterize truly general intelligence. Existing evaluation frameworks, which are centered on broad language or perception tasks, fail to capture generality at its core and offer no guidance. The artificial general intelligence testbed (AGITB) is a novel and freely available benchmarking suite comprising twelve fully automatable tests designed to evaluate low-level cognitive precursors through binary signal prediction. AGITB requires models to forecast temporal sequences without pretraining, symbolic manipulation, or semantic grounding. The framework isolates core computational invariants—such as determinism, sensitivity, and generalization—that align with principles of biological information processing. Engineered to resist brute-force and memorization-based approaches, AGITB presumes no prior knowledge and demands learning from first principles. While humans pass all tests, no current AI system has met the full AGITB criteria, underscoring its potential as a rigorous, interpretable, and actionable benchmark for guiding and evaluating progress toward artificial general intelligence. A reference implementation of AGITB is available on GitHub.

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

1 Introduction

Despite rapid advancements in machine learning and neural network architectures, artificial intelligence (AI) systems still fall short of flexible, robust 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 summarized why even the latest large reasoning models (LRMs), despite some impressive achievements, cannot be trusted. The surface-level competence obscures a deeper lack of understanding—an essential prerequisite for artificial general intelligence (AGI).

The progress in AI has led to 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 benchmarks 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 metric fully achieves its intended purpose; moreover, all lack key properties such as gradual resolution, interpretability, and full automatizability. They typically assess surface-level competence rather than underlying mechanisms of generalization. In response to these limitations, this paper introduces the artificial general intelligence testbed (AGITB), a novel framework for evaluating 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 twelve automatable, domain-agnostic tests. While AGITB does not aim to assess consciousness or semantic comprehension, it provides a principled mechanism for distinguishing narrow AI systems from those exhibiting generalizable, adaptive behavior. A comparative analysis with the Abstraction and reasoning corpus (ARC) introduced by Chollet (2019) is included to contextualize AGITB within the broader landscape of AGI evaluation.

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 often shifted. 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 in certain areas, we may inadvertently set a perpetually receding goalpost for AGI, failing to recognize 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 specialized benchmarks favor 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 specialized that humans have difficulty with them.

2.1 A Test That Humans Pass but Machines Fail

An effective AGI test should be intuitively solvable by 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 mem-

orization or domain-specific heuristics, requiring generalization, abstraction, and adaptive reasoning.

A valid AGI evaluation must either (1) reveal a fundamental cognitive gap between humans and machines or (2) define a behavioral 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 artificial general intelligence testbed 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 generalize patterns in raw binary signals independently of semantics or prior training. Rooted in the view—advanced by Hawkins and Blakeslee (2004)—that intelligence is fundamentally about extracting structure from data to make predictions, AGITB defines twelve core tests that conceptualize intelligence as a signal-based prediction. This approach aligns more closely with how biological intelligence appears to function at the cortical level—processing time-sensitive sensory data, not disembodied symbols.

3 Artificial General Intelligence Testbed

The primary objective of the artificial general intelligence testbed is to support the development and evaluation of AGI by defining a clear set of requirements that a model must meet to qualify as generally intelligent. A single model must fulfill the specific tasks defined in each test to validly claim success on the benchmark.

3.1 Components

AGITB requires the user to define two components: the `cortex`, representing the AGI system under evaluation, and the `input`, which encodes the data samples delivered to the cortex over time. The cortex is treated as a black box that predicts future inputs based on the observed history of prior signals.

Each input consists of binary-encoded data from (virtual) sensors and actuators. It is composed of a fixed number of bits, with each bit corresponding to a distinct input channel, such as an individual pixel, audio band, or actuator feedback signal. An input represents a snapshot of multiple parallel signals at a single point in time, as illustrated in Figure 1a. While spatial information is encoded within the structure of each input, semantic richness emerges through the temporal unfolding of input sequences, as shown in Figure 1b. Although the spatial and temporal dimensions are orthogonal, their interaction produces structured patterns to which the cortex must adapt.

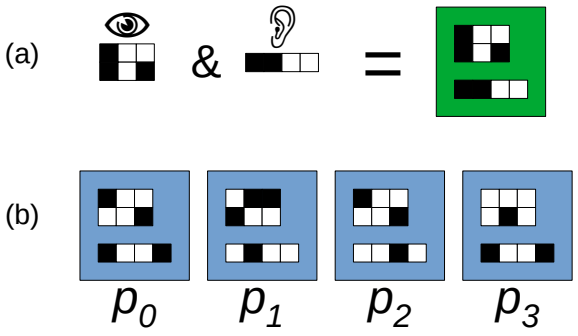

Figure 1: (a) Example of a 2×3-bit block of visual information combined with four bits of auditory data to form a single 10-bit input sample.
(b) Example of an input sequence illustrating a temporal pattern with a period of 4.

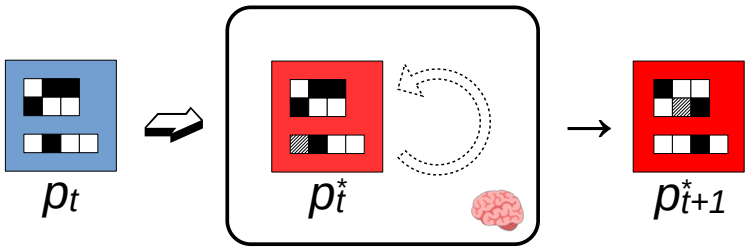

Figure 2: Adaptation to a sequence proceeds iteratively in discrete time steps. At the previous step $(t-1)$, the model predicted that the current input would be p^*_t; this prediction was incorrect in the first bit of the auditory data. Based on the actual input p_t, the model then generates a new prediction p^*_{t+1}, which is incorrect in the fifth bit of visual data for the sequence shown in Figure 1b.

3.2 Operation

The testbed presents the cortex with a stream of inputs over time. At each time step t, the cortex 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. In cases where predictions are not exact, proximity to the correct value is considered, which indicates the model's capacity for generalization and adaptation.

AGITB asserts expectations about the resulting state and behavioral dynamics of cortex models in specific scenarios using randomly generated test inputs. Rather than relying on arbitrary score thresholds, AGITB evaluates each model by comparing it against itself under controlled conditions. Since the internal state of the cortex is not accessible to the testbed for direct comparison, the user must define a custom criterion for determining model equality.

To execute AGITB, users must specify one parameter: the *pattern period*. This value determines the length of the repeating temporal sequence in the input stream that the cortex is expected to recognize and adapt to. Longer pattern periods increase the temporal complexity of the task, requiring more sophisticated memory and generalization abilities. However, when combined with high-dimensional and wide inputs, excessively long periods may surpass the model's learning capacity. As such, the pattern period should be selected in conjunction with input size to balance task difficulty with the model's capabilities, ensuring that each test remains feasible and diagnostically informative.

3.3 The 12 Essential Tests

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

Test 1 – Genesis

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 acquire understanding through experiential 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 anything other than an empty (spike-free) input before receiving any data is considered biased, as it encodes implicit assumptions about an unwarranted future. To maintain neutrality, such models must initialize to predict only empty patterns and remain unbiased until influenced by actual input.

Two cortex 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 cortex must possess an intrinsic organizational architecture capable of learning, even in the absence of prior data.

Test 2 – Bias

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

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

Test 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 cortex models must have experienced identical input histories. Any difference in input must lead to a divergence in state, reinforcing the principle that behavior 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.

Test 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.

Test 5 – Time

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

Because the cortex updates its state based on cumulative history, the order of inputs critically affects learning and adaptation. Recognizing and exploiting temporal structure is thus a defining feature of intelligent systems.

Test 6 – Refractory Period

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

Assertion: The cortex 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.

Test 7 – Temporal flexibility

Assertion: The cortex must be able to adapt to temporal patterns with the user-specified period.

Assertion: The cortex must also be capable of adapting to temporal patterns with periods longer than the user-specified value.

This test evaluates a model's capacity to generalize across temporal scales. Unlike rigid pattern-matching systems, intelligent models should flexibly detect and respond to recurring structures, regardless of their exact periodicity. The purpose of this test is to identify models that are overly specialized to a single temporal period rather than capable of robust generalization across multiple timescales.

Test 8 – Stagnation

Assertion: There exists a limit beyond which the cortex 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 cortex is exposed to input sequences that would otherwise be learnable.

Test 9 – Entropy

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

The structure of the input sequence affects how quickly a model can adapt. Simpler or more regular patterns typically lead to faster convergence, while complex or noisy inputs demand longer adjustment periods.

Test 10 – Salience

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

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

Test 11 – Unobservability

Assertion: Distinct cortices may exhibit the same observable behavior.

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

Test 12 – Generalization

Assertion: On average, adapted models achieve higher predictive accuracy than unadapted models after input disruption.

Assertion: On average, adapted models achieve higher predictive accuracy than random guessing after input disruption.

Only models capable of generalization will derive lasting benefit from prior learning when inputs are disrupted. Upon re-exposure to familiar stimuli, such models should, on average, outperform both unadapted models and random baselines in predictive tasks. Average performance is the evaluation criterion because even unadapted or random models may occasionally produce correct predictions by chance, without reflecting genuine understanding or learned structure.

3.4 Memorization

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

AGITB imposes a biologically inspired constraint: the refractory period, which prohibits any neuron (bit) from firing in consecutive time steps. This constraint significantly narrows the set of valid input sequences. The number of distinct length-n temporal input sequences of k bits, with the rule that a 1 never survives to the next time step, is $|S'| = (F_{n+2})^k$, where F_i denotes the i-th Fibonacci number, with $F_0 = 0$.

AGITB further reduces this space by requiring sequences to be cyclic, meaning that the first input must also satisfy the refractory constraint relative to the last input in the sequence. The number of distinct cyclic temporal input sequences with refractory constraint is given by $|S''| = (L_n)^k$, where L_i denotes the i-th Lucas number, with $L_0 = 2$.

Consider an illustrative example of a task with 3-bit inputs and a pattern period of $n = 5$. The total number of unconstrained sequences is $|S| = 2^{15} = 32.768$. Applying the refractory constraint reduces the valid set to $|S'| = 13^3 = 2.197$ sequences. When the cyclic condition is added, the space shrinks further to $|S''| = 11^3 = 1331$.

Even with these biologically grounded constraints, the size of the valid sequence space grows exponentially concerning both sequence length and input dimensionality. Given that real-world sensory inputs may span millions of bits, strategies based on brute-force memorization are computationally intractable. The exponential complexity highlights the necessity for models that rely on generalizable, pattern-based learning mechanisms rather than rote memorization.

3.5 Performance

Before assessing AGITB's utility, it is important to recognize its role as a pragmatic benchmark, not as an end goal in itself. Like the Turing Test, which Harnad (1992) described as an empirical criterion rather than a philosophical sleight of hand, AGITB is intended to

serve as a practical tool for evaluating progress toward AGI. The ultimate objective remains AGI's development, not merely passing the test.

AGITB can yield meaningful insights only when developers adhere faithfully to its core requirements. Misinterpreting elements such as the "empty" initial state can lead to unreliable or misleading results.

AGITB provides a structured framework for empirically evaluating the AGI potential of various systems, including classical symbolic architectures, artificial neural networks (ANNs), and large language models. However, before benchmarking artificial systems, it is first necessary to consider the baseline performance of human cognition.

3.5.1 Human Performance

The inability to compare internal cortical states makes it impossible to directly verify AGITB tests in humans. However, because the cortical architecture naturally supports low-level binary signal processing and the tests align with cognitive expectations, we can assume that humans inherently satisfy most of the 12 tests. Tests 1 (Genesis) and 6 (Refractory Period), however, require additional argumentation.

Due to prior experience and cognitive bias, an adult human may appear to "fail" the first test (Genesis), since their cortex is no longer in a truly unconditioned state and may produce a non-empty prediction. In contrast, a fetal cortex—with no prior synaptic structure—receiving its first-ever input would satisfy the condition of true neutrality by predicting no spikes.

AGITB includes the refractory period requirement (Test 6) because it reflects biological reality; however, it does not eliminate the possibility of a deviating solution. Such a solution could feign compliance with the requirement without genuinely supporting it, whereas the converse is not possible.

Additionally, the more complex tests have counterparts at the cognitive level, where their effects can be observed through conscious reasoning and introspection. Temporal flexibility (Test 7), for example, is simple for humans as we have no problems recognizing temporal patterns of varying durations. The property of stagnation (Test 8) is a consequence of a (human) cortex being physically limited in its capacity to store knowledge. The test checks for the first sign of dementia—a moment when the cortex becomes saturated and forgets the recent event. Because humans learn different topics at varying speeds, the rationale for the entropy (Test 9) and the salience requirement (Test 10) is self-evident. The inability to directly inspect human mental states is captured in the unobservability requirement (Test 11). Finally, the generalization test (Test 12) requires what humans naturally excel at: generalizing when faced with new but distorted input.

3.5.2 Classical Symbolic Programs

AGI approaches rooted in classical programming are fundamentally constrained by the Symbol Grounding Problem (SGP), as explained by Harnad (1990). While such systems may exhibit weak AI capabilities—producing behavior that appears intelligent—they lack grounded understanding and, therefore, cannot be considered genuine instances of AGI, as argued by Searle (1980).

In classical AI systems, a model's initial state is inseparable from its program, which encodes fixed logic and predetermined responses. Consequently, such systems cannot begin from an unbiased or knowledge-free state, as their behavior is entirely dictated by human-authored instructions. AGITB's first test, however, introduces a paradox: the system must initialize without any such program—that is, it must begin in a truly neutral state, free from prior knowledge or assumptions. If they adhere to the empty-state requirement, they remain inert; if they do not, they reintroduce the symbol grounding problem. In the language of Searle's Chinese Room Argument, the ledger must be blank—containing no predefined instructions for interpreting Chinese symbols. Yet classical systems require a program to function; absent such instructions, they halt.

AGITB's initial test formalizes the requirement to learn from scratch: it metaphorically demands that Searle acquire a language—such as Chinese—purely through exposure to input, without access to any prior rules. To enforce this condition reliably, the model comparison must be rigorous, ensuring that all relevant internal properties of the system are examined. Any structural discrepancy or hidden initialization could conceal bias and compromise the integrity of the test.

It is essential that models begin from a genuinely unbiased state. As an external framework, AGITB cannot independently verify whether a program that declares itself "empty" is truly devoid of prior knowledge. Any attempt to obscure bias at initialization compromises the integrity of the benchmark and invalidates the results of subsequent tests. By embedding prior assumptions, the developer effectively undermines the model's claim to general, input-driven learning.

3.5.3 Artificial Neural Network Performance

Unlike classical symbolic programs, artificial neural networks do not encounter the same initialization paradox. Perceptron-based architectures are inherently robust: they always remain operational, as neurons continuously perform computations, even when their input remains neutral (for example, when an activation function returns zero).

However, the requirement that a model begins in a completely unbiased, knowledge-free state stands in fundamental tension with the prevailing paradigm of modern deep learning. Contemporary approaches typically rely on pretraining, which adjusts the network's weights in advance through exposure to structured or labeled data. What remains absent is a mechanism by which an uninitialized network can autonomously begin learning—without supervision, without predefined semantics, and without relying on external scaffolding.

Although an untrained network may technically operate without internal excitation, no learning dynamics are triggered. The neurons are not halted but dormant and passively awaiting informative input. This reveals a critical limitation of current ANN designs: they lack the intrinsic capability to initiate unbiased, autonomous learning from scratch in arbitrary, unfamiliar environments.

3.5.4 Large Language Model Performance

Because LLM is essentially an ANN, it fails the genesis test. This failure immediately indicates that LLMs cannot learn entirely on their own and are inherently biased toward the language and data they were trained on. Although the internal state of an LLM is

readily accessible and could, in principle, be used to perform other tests, such evaluations are meaningless until the first requirement is satisfied.

Consequently, LLMs cannot acquire new knowledge—let alone learn a new language—solely from prompts, nor can they develop the kind of grounded understanding of the world that humans possess. This implication also applies to large reasoning models, which are themselves built upon LLM architectures.

3.6 Remarks

AGITB evaluates a 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 minimize 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.

This 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. AGITB's "all-tests-must-pass" standard enforces a rigorous, biologically motivated foundation for evaluating AGI, grounded in the same principles that underlie natural intelligence.

While individual AGITB tests may seem trivial in isolation, the central challenge lies in designing a universal AGI architecture capable of satisfying 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 symbolic system, and it remains an open question whether such a solution is even theoretically possible. If a purely symbolic system were to succeed, it would suggest that AGITB fails to effectively differentiate narrow AI from genuinely general intelligence, given that symbolic systems—while proficient in syntactic manipulation—arguably lack true semantic understanding, as discussed by Searle (1980). Until such a system is demonstrated, AGITB's assertions may be regarded as necessary (though not sufficient) conditions for identifying genuine, domain-independent intelligence. Although passing AGITB does not equate to achieving full AGI—since it does not, for instance, assess social reasoning or natural language fluency—failure to meet its conditions likely indicates a deficiency in foundational general cognitive capabilities.

Hand-engineered, task-specific systems have historically struggled to scale toward true generalization, reinforcing the need for adaptive, learning-based architectures such as artificial neural networks. However, state-of-the-art 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 framework?

3.7 Cheating the Benchmark

Since AGITB tests are not inherently difficult to solve in isolation, a plausible strategy for circumventing the benchmark would be to develop task-specific models and deploy them selectively based on the task at hand.

While it is theoretically possible to infer a 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 Cortex programming interface—these measures would not only deter cheating but also risk making tests less intelligible to human developers. This would undermine two of the testbed's core objectives: transparency and interpretability.

For this reason, the AGITB reference implementation remains deliberately readable and accessible. To date, no classical symbolic AI system has demonstrated the flexibility or generality required to handle the arbitrary patterns of structure and length 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 AGITB and ARC: A Comparison

The absence of a unified definition of intelligence has led to AGI benchmarks that often suffer from conceptual ambiguity and limited practical utility. Most offer little insight into the mechanisms underlying generalization. In this regard, AGITB takes a distinct approach. Focusing on the emergence of intelligence from low-level signal-processing building blocks offers a biologically grounded and mechanistically interpretable framework.

Among competing benchmarks, the Abstraction and Reasoning Corpus (ARC), introduced by Chollet (2019), is perhaps most aligned in spirit, as it also prioritizes generalization over task-specific optimization. ARC presents visual reasoning tasks that require a model to infer novel transformations from a small set of input-output examples—such as recoloring, 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 neocortex—the primary focus of the benchmark—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 recognize 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 presupposes intelligence, whereas AGITB requires it to emerge from first principles. ARC evaluates performance based on behaviors grounded in human-like cognitive priors, while AGITB measures a system's ability to develop necessary priors autonomously 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 conceptualizes intelligence as a universal capacity for learning independent of innate symbolic structures or species-specific assumptions. Table 1 outlines the key differences between the two benchmarks.

Property	ARC	AGITB
Learning mode	Online	Online
Susceptible to memorisation	No	No
Interface modality	Visual	Binary
Target intelligence type	Human	Universal
Assumed cognitive priors	Yes	No
Model development feedback	No	Yes
Fully automatable	Limited	Yes
Task type	Synthetic	Synthetic
Task production	Manual	Automatic
Task inspiration	High-level reasoning	Low-level processing
Task spatial dimensions	2	∞
Task temporal structure	No	Yes

Table 1: Comparison of core properties between the ARC and AGITB benchmarks.

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 behaviors 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 generalization 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.

Acknowledgments and Disclosure of Funding

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

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.

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.

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.

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.