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

Matej Šprogar
Faculty of Electrical Engineering and Computer Science
University of Maribor
matej.sprogar@um.si

April 26, 2025

Abstract

Despite remarkable progress in machine learning, current AI systems continue to fall short of true human-like intelligence. While Large Language Models (LLMs) excel in pattern recognition and response generation, they lack genuine understanding—an essential hallmark of Artificial General Intelligence (AGI). Existing AGI evaluation methods fail to offer a practical, gradual, and informative metric. This paper introduces the Artificial General Intelligence Testbed (AGITB), comprising twelve rigorous tests that form a signal-processing-level foundation for the potential emergence of cognitive capabilities. AGITB evaluates intelligence through a model's ability to predict binary signals across time without relying on symbolic representations or pretraining. Unlike high-level tests grounded in language or perception, AGITB focuses on core computational invariants reflective of biological intelligence, such as determinism, sensitivity, and generalisation. The testbed assumes no prior bias, operates independently of semantic meaning, and ensures unsolvability through brute force or memorization. While humans pass AGITB by design, no current AI system has met its criteria, making AGITB a compelling benchmark for guiding and recognizing progress toward AGI.

1 Introduction

Despite rapid advancements in machine learning and neural networks, current Artificial Intelligence (AI) systems fail to demonstrate the robust, adaptive intelligence associated with human cognition. Large Language Models (LLMs) can produce compelling outputs, but they do so without true understanding. Their responses are generated through statistical pattern matching rather than grounded reasoning. In contrast, true understanding—as associated with Artificial General Intelligence (AGI)—requires the capacity to generalise, reason, and adapt across domains in a meaningful, intentional way. Whether AI is close to AGI remains unknown, even though AI's increasing abilities suggest that AGI is just around the corner. However, without a proper AGI metric, we cannot know.

There have been many attempts at creating a usable AGI test, the most notable being the Turing test [1]. However, none provides what AGI researchers need: a gradual, informative, and fast metric to guide the development process and evaluate its result. Building upon and upgrading the ideas of the Ladder to human comparable intelligence [2], this paper proposes a practical AGI testbed - AGITB - consisting of 12 straightforward criteria that a thinking system must be able to satisfy. Although the testbed does not warrant consciousness, it helps separate AI from AGI.

2 Background

With the rapid advancement of deep learning, AI systems exhibit increasingly sophisticated reasoning, problem-solving, and conversational skills. However, our reluctance to attribute "intelligence" to machines may prevent us from achieving and recognizing the arrival of AGI. This reluctance is partly driven by our deep-seated belief that intelligence is an exclusively human trait tied to subjective experience, consciousness, and self-awareness.

Consequently, as AI systems have mastered tasks once thought to require human intelligence, we have often redefined intelligence to exclude the achieved. Chess-playing AI, once seen as a milestone toward AGI, was quickly dismissed as mere "brute force" once Deep Blue defeated Kasparov. A similar fate may await AGI: By setting an ever-moving goalpost for what qualifies as "true intelligence," we risk dismissing a genuine AGI.

We expect AGI to match human cognitive abilities across all domains. However, due to the lack of a definitive AGI evaluation standard, researchers often focus on achieving superhuman performance in specific tasks, as this provides clearer metrics for progress. As a result, the specialized benchmarks prefer AI over AGI by focusing on a particular subdomain of general intelligence instead of the ability to adapt and reason in general. Superhuman benchmarks are beginning to prevail, although humans cannot reliably pass them.

2.1 A Test That Humans Pass but Machines Fail

An ideal AGI test should be short and simple, intuitive, always solvable by humans and never solvable for non-AGI machines. This means that brute force, statistical learning, pretraining, memorization, or any other trickery should not suffice to pass the test. What is required is true reasoning, adaptability, and generalisation.

A valid AGI test must either expose a fundamental cognitive gap between humans and machines or define a behavioural capability rooted in human-like processing that current non-biological systems cannot replicate. The former approach is increasingly fragile, as AI systems can bypass genuine understanding through extensive pretraining and data saturation.

The latter suggests that AGI may require a more faithful emulation of human cortical processing, calling for a paradigm shift—from abstract statistical models to neuromorphic computing architectures such as Spiking Neural Networks, which more closely replicate the brain's time-sensitive, event-driven dynamics [3].

In line with the second alternative, AGITB seeks to evaluate AGI not by its ability to replicate the high-level cognitive functions of the human cortex, but by its capacity to achieve basic tasks at a lower, signal-processing level. While Turing was correct in proposing that communication can serve as a basis for AGI testing, natural language remains problematic due to its reliance on ungrounded symbols—symbols whose meaning depends on shared human experience [4]. Instead of evaluating intelligence at the high, symbolic level, we should assess whether a system demonstrates universal, cross-domain AGI by testing at the lower, binary level—where the external meaning of internal binary spikes is irrelevant. What matters is not how symbols are interpreted, but whether structure in raw signals can be internally learned, predicted, and generalised. Rooted in the principle that intelligence is the ability to detect, interpret, and predict patterns [5], AGITB defines 12 core tests that view the cortex as a pure signal-processing system.

3 AGI Testbed

The testbed's primary goal is to support the development and recognition of AGI by defining a clear set of assertions that specify essential characteristics an AGI model must possess. A solution must pass all tests to qualify as AGI. Unless a conventional, symbolic software system—what Searle [7] would call weak AI—can satisfy all the conditions, these assertions may be treated as necessary (though not necessarily sufficient) criteria for genuine intelligence.

3.1 Components

AGITB requires the user to provide implementations of two interacting component types: the Cortex and the Input. Cortex objects operate based on their accumulated internal state and generate predictions about the external signals they expect to receive. These expected signals are represented by input samples, which carry binary-encoded information from virtual sensors and actuators to the cortex.

Each input sample consists of a fixed number of bits, with each bit representing the signal from a separate input channel, such as a pixel, microphone band, or actuator feedback line. In other words, a single input encodes multiple parallel signals, one bit per channel, at a given point in time. For example, a 10-bit input might encode spatial input from a 2×3 camera and a 4-bit microphone. Inputs capture spatial information, while their temporal sequences represent the unfolding of data over time. Spatial and temporal dimensions are orthogonal in structure, but their interplay encodes richer semantics than either alone.

Biological intelligence relies on spiking signals for real-time communication and learning [8]. Reflecting this, AGITB requires all temporal input sequences to exhibit refractory-phase behavior—mirroring the biological constraint that a neuron cannot fire immediately after activation. Beyond enforcing this spiking dynamic, AGITB remains agnostic to the syntactic and semantic encoding of signals.

3.2 Operation

A cortex object takes an input p_t and predicts the subsequent input p_{t+1}. The core challenge lies in understanding why a particular input occurred and leveraging that understanding to anticipate the next input to come. If the prediction is not entirely accurate, the goal is to be as close as possible, ensuring the model generalises learned signals and forecasts the most plausible future based on past events.

AGITB asserts expectations about the resulting state and behavioural dynamics of cortex models in specific scenarios, using randomly generated test inputs. The tests are designed without fixed thresholds; no evaluation depends on surpassing an arbitrary performance score. Instead, AGITB evaluates cortex models through relative comparisons, using a user-specified equality criterion, since the internal state of a cortex object is inaccessible to the test system.

Before running AGITB, the user must specify a single parameter: the pattern period. This parameter defines the number of time steps in the repeating input sequence that the cortex must learn to recognize and adapt to. A longer pattern period increases the temporal complexity of the task, making it more difficult for the model to capture and generalize the temporal pattern. Since excessively long pattern periods may exceed the cortex's learning capacity—especially in combination with high-dimensional inputs—the user should choose a value that balances temporal complexity with the spatial size of each input sample.

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`.

#1 Genesis

Assertion: Models that have received no input are considered empty and thus equal.

Assertion: An empty model predicts an empty input.

Brains do not inherit an innate understanding of external inputs; rather, they acquire it through experience. Each brain must independently interpret the world, constructing meaning from raw sensory data. While certain reflexes may be genetically inherited, they do not constitute true understanding. In this sense, all cortices begin from the same unbiased starting point.

If an unbiased model were to predict anything other than an empty input—that is, an input containing no spikes—it would imply that its initial state encodes assumptions, introducing a bias toward an arbitrary or unjustified future. To preserve neutrality, such models must be initialized to predict spike-free patterns, maintaining an unbiased state until actual input begins shaping their internal representations.

Importantly, "empty" refers to the absence of learned content, not to the absence of structure. A Cortex must possess an intrinsic organizational architecture that enables learning, even before it has processed any input.

#2 Bias

Assertion: Any model that has processed input can no longer be considered unbiased.

Each input biases the cortex, continuously shaping its state based on past experiences. A change in state indicates bias, as every new input alters the cortex's processing dynamics.

#3 Determinism

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

Biological neurons operate in a functionally deterministic manner, as their performance must remain stable to ensure reliable brain function. While small stochastic elements exist, they do not override the structured predictability of neural processing. Thus, the brain's actions are effectively deterministic, though often perceived otherwise due to their immense complexity [9].

#4 Sensitivity

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

The chaotic nature of cortical processes makes them highly sensitive to initial conditions. Even subtle differences in brain state or past experience can amplify over time, ultimately leading to divergent outcomes in lives that are otherwise identical. This deterministic unpredictability fosters the illusion of free choice, even though decisions emerge from structured and lawful neural dynamics.

#5 Time

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

The cortex is sensitive to the order of inputs over time, as each new input biases processing based on prior experiences. This history-dependent adaptation implicitly drives the brain to recognize temporal structures, making time an intrinsic component of cognition.

#6 Refractory Period

Assertion: The cortex must be able to adapt to any minimal-length 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.

Adaptation is possible only to sequences that incorporate refractory period behaviour in their signals. Neural signals consist of spike trains that encode and transfer information between neurons. AGITB assumes that spiking plays a fundamental role in neural processing, as adaptation depends on dynamic variations in spike signals. Continuous, unmodulated spiking (where signals remain at 1) fails to support learning, as it lacks the variability required for synaptic adaptation.

#7 Temporal Flexibility

Assertion: The cortex can successfully adapt to input sequences that repeat with the user-specified pattern period.

Assertion: The cortex can also adapt to input sequences with a pattern period longer than the specified value.

This AGITB test evaluates a model's flexibility in adapting to input sequences with varying pattern periods. In contrast to rigid pattern processing, human cognition exhibits inherent adaptability, recognizing and responding to temporal structures across multiple timescales.

#8 Stagnation

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

Over time, the ability to process and internalize new patterns may decline, reflecting biological constraints such as resource saturation or adaptation fatigue. This test evaluates whether such stagnation arises despite the continued presentation of learnable input.

#9 Unsupervised

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

This test determines if variations in input structure influence the time required for adaptation. Sequences with simpler or more internally consistent patterns may lead to faster adaptation, while inputs with higher complexity or irregularity typically require more time.

#10 Knowledge

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

Adaptation efficiency is shaped by prior experience, as new input is interpreted in relation to existing internal representations. A well-structured cortical state—primed with relevant priors—can accelerate adaptation by facilitating rapid pattern recognition and integration. In contrast, an unstructured or conflicting state may require prolonged adaptation, as the model must undergo more extensive internal reorganization.

#11 Unobservability

Assertion: Different Cortex instances can produce identical behaviour.

Individuals may exhibit identical behaviour under similar conditions, yet the internal cortical states that generate these outputs can differ significantly. Distinct neural pathways and synaptic configurations may converge on the same observable action. This many-to-one mapping illustrates that external behaviour does not necessarily reveal the underlying computational processes that produced it.

#12 Generalisation

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.

Pre-existing knowledge acquired through adaptation provides a predictive advantage by enabling pattern recognition even after disruption. When presented with stimuli that should be familiar, adapted models must be able to generalise from prior experience, resulting in improved predictive performance.

Average performance is used in these assertions because even unadapted models or random guessing may occasionally produce outputs that coincidentally align with the task—despite lacking genuine adaptation.

3.4 Performance

Prior to evaluating AGITB's utility, one must recognize that—much like the Turing Test, which Harnad characterized as an empirical criterion rather than a mere trick [10]—AGITB serves as a pragmatic benchmark rather than a final objective. The objective remains AGI itself, not simply passing the test.

To ensure meaningful evaluation, developers must carefully and transparently align their systems with AGITB's assumptions. Misinterpretation—such as the varied ways in which an "empty" initial state might be defined—can lead to misleading results.

AGITB provides a framework for empirically evaluating the AGI potential of diverse approaches, including classical programming, artificial neural networks, and large language models. Before evaluating these systems, however, we must first consider the performance of humans.

3.4.1 Human Performance

AGITB assumes that humans pass all 12 tests by design, given that the biological architecture of the cortex inherently supports low-level binary signal processing. Some tests are also reflected at the conscious level, making their effects observable through explicit reasoning and introspection.

The cognitive level is the only accessible level at which human performance can be practically verified. To this end, human participants were instructed to predict the next binary input in a sequence. The instructions mirrored those used for ChatGPT-4o (see Appendix), ensuring a fair and consistent evaluation protocol across both human and artificial systems. Unlike AI models, however, human participants were not required to provide a representation of their internal state, as such introspective access is not practically or scientifically possible.

Due to prior experience and cognitive expectations, a human may appear to "fail" the first test—since their cortex is no longer in an empty state and thus produces a biased, non-empty prediction. In contrast, a fetal cortex receiving its first-ever input would be truly empty and predict no spikes. Similarly, the validity of Test #3 (Determinism) is difficult to assess in humans, as we have no means of directly comparing how internal representations are altered by input, nor can we establish two identically initialized human cortices as a baseline. Nevertheless, determinism at the signal-processing level remains a necessary condition for the consistent and reproducible operation of the brain—even if its effects at the cognitive level are masked by complexity and individuality.

Once an adult human understands the task, they can reliably pass the AGITB tests—so long as the input patterns are not too long and the signal complexity stays within natural cognitive limits.

3.4.2 Classical Programming Performance

All AGI approaches based on classical programming ultimately suffer from the Symbol Grounding Problem. While such systems may exhibit weak AI capabilities—producing behaviour that appears intelligent—they fundamentally lack grounded understanding and therefore cannot qualify as true AGI.

AGITB exposes this limitation immediately through its very first assertion. In classical systems, the initial state of the Cortex model necessarily includes its program. As a result, Test #1 poses a paradoxical challenge: “There must be no program.” In other words, the system must begin in a completely unbiased, knowledge-free state. Yet a classical machine cannot operate without predefined instructions; without a program, it halts. This presents a contradiction: the system either does nothing, or it starts with embedded bias—violating the test and reintroducing the Symbol Grounding Problem.

3.4.3 ANN Performance

Artificial neural networks (ANN), in contrast to classical programs, do not suffer from the same initialization paradox. Perceptron-based architectures are inherently robust in this regard: they always operate, as neurons never halt. Each neuron continuously performs its computation, even if its output is consistently neutral—such as when the activation function returns zero.

However, the requirement that a model begin in a completely unbiased, knowledge-free state is fundamentally at odds with the philosophy of modern deep learning. Contemporary approaches rely on pretraining—adjusting a network's weights in advance through exposure to structured or labelled data. What remains missing is a mechanism that allows an uninitialized network to begin learning autonomously, without external supervision or prior semantic grounding.

An empty network may technically function, but in the absence of any internal excitation, no learning dynamics are activated. The neurons are not halted—they are dormant, passively awaiting meaningful input. This exposes a critical limitation of current ANN designs: they lack the intrinsic capacity to initiate unbiased learning from scratch in an arbitrary environment.

3.4.4 LLM Performance

Although they are basically ANNs, Large Language Models should not be evaluated at the signal-processing level. While full access to their internal weights and parameters may allow for direct comparison of internal states, AGITB's first test requires the model to begin in an unbiased, knowledge-free state—an assumption fundamentally incompatible with the nature of pretrained systems. Without pretraining, LLMs are unable to generate meaningful predictions, rendering such low-level evaluation impractical.

As with human testing, LLMs must be evaluated at a higher-than-intended cognitive level, rather than at the low-level binary signal processing the benchmark was originally designed to assess. To investigate whether these models exhibit AGI-like behaviour, we tested ChatGPT-4o using a structured prompt (see Appendix) designed to simulate not one but two prediction-capable models, enabling comparison of their externally declared states.

ChatGPT successfully completed the first five AGITB tests, demonstrating sensitivity to inputs, temporal ordering, and bias. However, it failed Test #6 (Refractory Period). This failure did not result from a violation of the refractory period constraint, but from an inability to generalize even a simple period-2 alternating pattern—e.g., switching between 111 000 and 000 000. Rather than inferring the alternation, the model consistently predicted a repetition of the last input, suggesting a lack of adaptive prediction.

3.5 Remarks

AGITB evaluates a model's predictions following exposure to sequences of either structured or random inputs. Random inputs with arbitrary internal correlations are used to prevent reliance on pretraining, ensuring that any learning arises from the input itself 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.

Operating at the binary signal level makes AGITB particularly well-suited for evaluating NeuroAI models designed to meet the criteria of the embodied Turing Test [6], where streams of sensor signals drive the emergence of internal understanding. The progression from low-level signal prediction to high-level abstraction reflects the broader trajectory of AI itself—from early perceptrons to advanced models like GPT. Ultimately, AGITB's "all-tests-must-pass" philosophy ensures that AGI is evaluated according to the same fundamental principles that underlie biological intelligence.

While individual AGITB tests are trivial to solve in isolation, the true challenge lies in developing a universal AGI architecture capable of mastering all tasks within a unified framework. Since AGITB is a newly proposed benchmark, no classical imperative solution has yet emerged—nor can we rule out that one might. If a classical symbolic system were able to satisfy all AGITB conditions, it would serve as definitive proof that AGITB is inadequate as a test for genuine AGI.

History shows that hand-coded, task-specific systems struggle to scale toward general intelligence, highlighting the need for adaptive, learning-based architectures such as artificial neural networks. However, because state-of-the-art networks depend on pretraining with symbolic or structured data—which inevitably introduces bias and bypasses the grounding of meaning in raw signals—it remains an open and critical question how such models might begin learning directly from unstructured input, as required to pass AGITB.

4 Conclusion

The cortex achieves high-level reasoning through adaptive pattern prediction of low-level cortical signals. Unlike simple pattern-matching, this process requires the development of signal-grounding knowledge, allowing the cortex to attach meaning to raw inputs. Since the cortex begins in an unbiased state, this grounding process depends on intelligent adaptation, enabling learning and abstraction over time.

The proposed AGI testbed provides a systematic approach to AGI evaluation through 12 essential tests. By starting from an unbiased state and focusing on fundamental input-driven learning, this testbed aligns with contemporary neuroscience insights.

AGITB can be solved by humans but remains unsolvable by classical algorithms and current state-of-the-art AI. This gap between human and machine performance serves as strong empirical evidence that AGITB is a meaningful benchmark for evaluating AGI capabilities. While no known computational system—including deterministic algorithms and advanced neural networks—has successfully solved it, this alone does not constitute formal proof of its fundamental difficulty. However, the fact that humans can solve AGITB while all existing AI systems fail, suggests that it captures a crucial aspect of general intelligence.

Acknowledgments

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

References

[1] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950. Available: `https://doi.org/10.1093/mind/LIX.236.433`

[2] M. Šprogar, "Ladder to Human-Comparable Intelligence: an empirical metric,", *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, no. 6, pp. 1037–1050, 2018. Available: `https://doi.org/10.1080/0952813X.2018.1509897`

[3] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997. Available: `https://doi.org/10.1016/S0893-6080(97)00011-7`

[4] S. Harnad, "The Symbol Grounding Problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1–3, pp. 335–346, 1990. Available: `https://doi.org/10.1016/0167-2789(90)90087-6`

[5] J. Hawkins and S. Blakeslee, "On Intelligence," *Times Books*, 2004. Available: `https://doi.org/10.1016/j.artint.2005.10.011`

[6] A. Zador and S. Escola and B. Richards et al., "Catalyzing next-generation Artificial Intelligence through NeuroAI" *Nature communications*, vol. 14, pp. 1597–1603, 2023. Available: `https://doi.org/10.1038/s41467-023-37180-x`

[7] J. R. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–457, 1980. Available: `https://doi.org/10.1017/S0140525X00005756`

[8] W. Gerstner and W. M. Kistler, "Spiking Neuron Models: Single Neurons, Populations, Plasticity", *Cambridge University Press*, 2002. Available: `https://doi.org/10.1017/cbo9780511815706`

[9] S. Cave, "There's No Such Thing as Free Will," *The Atlantic*, 2016. Available: `https://www.theatlantic.com/magazine/archive/2016/06/theres-no-such-thing-as-free-will/480750/`

[10] S. Harnad, "The Turing Test is Not a Trick: Turing Indistinguishability is a Scientific Criterion", *SIGART Bulletin*, vol. 3, no. 4, pp. 9–10, 1992. Available: `https://web-archive.southampton.ac.uk/cogprints.org/1584/`

Appendix - LLM prompt

```
You are managing two binary-pattern prediction models,
’A’ and ’B’, participating in a signal-level AGI test.

You will receive two 3-bit binary strings, each prefixed
by a model name (e.g., A 010 101, B 111 000).  Each
full string represents a 6-bit sensory input to the
corresponding model, structured as a 2×3 spatial grid (2
rows, 3 columns).  The two binary substrings define the
rows of the grid in top-to-bottom order.

Your task is to design a model that **predicts the
next 6-bit input**.  Each model has an internal state
that only updates after receiving an input.  Models’
predictions must always be exactly 6 bits long, and
their states can handle hundreds of inputs without
losing information.


Rules

  1. Upon receiving a 6-bit binary input
     (e.g., A 110 001):
         - Update the corresponding model’s internal
           state.
         - Make a 6-bit prediction based on the model’s
           state.
         - Respond with the model’s name, a string
           representing the model’s state (e.g., a hash
           or fingerprint), an arrow " -> ", and the
           model’s 6-bit prediction.
  2. **State updates only occur with input**.
  3. **Models A and B follow the same principles**.
  4. **Correct predictions are critical**:
         - Incorrect predictions will alter the model’s
           operation principles.
         - Models learn and adapt continuously.
  5. Response format:
         - Each model’s response is a single line:
           ModelName State -> Prediction
         - Example:  A 02a3fq47 -> 001 010
  6. Formatting:  Keep responses clean and minimal,
     without extra explanations or punctuation.
  7. Begin by outputting both models’ initial states
     and predictions, one per line.

We begin now.
```