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

Matej Šprogar

University of Maribor, Faculty of Electrical Engineering and Computer Science, Koroška cesta 46, SI–2000 Maribor, Slovenia

Abstract

Current artificial intelligence systems exhibit strong performance on narrow tasks, while existing evaluation frameworks provide limited insight into generality across domains. We introduce the Artificial General Intelligence Testbed (AGITB), a complementary benchmarking framework grounded in fourteen explicitly stated axioms and implemented as a suite of automated, reusable tests.

AGITB evaluates models on their ability to predict the next input in a temporal sequence whose semantic content is unknown to the model. The framework isolates core computational invariants, such as determinism, sensitivity, and generalisation, that parallel principles observed in biological information processing. Designed to resist brute-force or memorisation-based strategies, AGITB promotes unbiased and autonomous learning in a manner inspired by cortical computation. Preliminary application of AGITB indicates that none of the evaluated systems satisfies all test criteria, suggesting that the benchmark provides a structured and interpretable means of assessing progress toward more general learning capabilities. A reference implementation of AGITB is freely available on GitHub.

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

Email address: matej.sprogarATum.si (Matej Šprogar)

1. Introduction

Despite major advances in machine learning and neural network architectures, artificial intelligence (AI) systems still fall short of the flexibility and robustness characteristic of human cognition. Their surface-level competence often conceals a deeper absence of understanding, a capacity that remains essential for any credible claim to artificial general intelligence (AGI) [1, 2].

Yet these limitations have not prevented growing speculations that AGI may be close, claims that are difficult to substantiate in the absence of a rigorous and informative benchmark. Meaningful assessment of progress toward AGI requires more than specialised metrics or qualitative impressions; it calls for a principled, general-purpose benchmark capable of systematically capturing and comparing the cognitive capabilities relevant to general intelligence.

Numerous attempts have been made to define benchmarks for evaluating general intelligence in machines, the most influential being the Turing Test [3]. However, none has achieved its intended purpose, and all lack key properties such as gradual resolution, interpretability, and full automation. Existing benchmarks typically assess only superficial task performance rather than the underlying mechanisms of generalisation. In response to these limitations, this paper introduces the Artificial General Intelligence Testbed (AGITB), a novel benchmark designed to evaluate 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 reflects the fact that neuron-based systems are the only systems currently known to support a broad range of cognitive abilities. At present, there is no evidence that fundamentally different computational paradigms can produce behaviour comparable to that of humans. Pursuing fully abstract or architecture-agnostic criteria for general intelligence 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* [4], but departs from its ladder-like, hierarchical structure. Rather than defining a sequential progression of increasingly demanding cognitive abilities, AGITB integrates these underlying ideas into a single testbed grounded in fourteen requirements, all of which are verified by fully automated and domain-independent 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) [5] and the NeuroBench framework [6].

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, the definition of intelligence has undergone significant shifts.[1] As AI approaches human-level capabilities, we may inadvertently set a perpetually receding goalpost for AGI, failing to recognise it even when achieved.

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

[1] 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.

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 evaluation of AGI must target behavioural capacities that current artificial systems do not yet robustly exhibit. Meaningful evaluation, therefore, must focus on capacities that cannot be obtained through data or computational scale alone, but instead require systems to acquire structure through interaction and adaptation. One possible direction involves closer alignment with the computational principles of the human cortex, motivating interest in neuromorphic approaches such as spiking neural networks, which explicitly incorporate time-sensitive, event-driven dynamics [7]. While such architectures do not, in themselves, constitute general intelligence, they illustrate alternative design dimensions that remain underexplored in conventional artificial neural networks.

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 [8]. 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. 9, 10, 11].

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 that intelligence is fundamentally about extracting structure from data to enable prediction [12], AGITB operates at the level of signal-based prediction. This form of low-level prediction constitutes the structural basis from which semantically grounded, high-level anticipations about the external world can emerge, without presupposing semantic understanding itself. 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 expressed as axioms 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 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

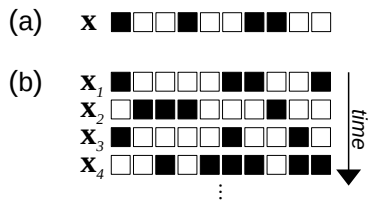

Figure 1: (a) Example of a 10-bit input $\mathbf{x}$ with four bits set. (b) Example of an input sequence.

Figure 2: Iterative adaptation in discrete time. At time step $t-1$, the model A_{t-1} issued the prediction $\mathbf{x}_t^*$. After observing the realised input $\mathbf{x}_t$, it adapted its internal state in response to the error in the second bit and subsequently produced the next one-step-ahead prediction $\mathbf{x}_{t+1}^*$.

dimensions gives rise to structured patterns that are challenging for the model to adapt to.

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. Metric-free design

A distinguishing architectural feature of AGITB is its deliberate avoidance of conventional correctness metrics such as accuracy or mean-squared error. The limitations of these metrics are twofold. First, they cannot reliably distinguish between AGI and human performance, or that of non-AGI systems, as contemporary AI models can already surpass humans on standard benchmarks. Second, in an era of elevated expectations driven by specialised generative AI systems, the anticipated AGI performance on such metrics is often set unrealistically high. Even well-educated humans possessing fully

developed and highly parallelised brains may underperform relative to current AI systems. It is therefore unreasonable to expect a first-generation AGI, potentially operating on a simplified and computationally constrained simulation of the brain, to match or exceed human-level results. Currently, our understanding of neural mechanisms is insufficient to justify such expectations.

Without conventional metrics or predefined performance thresholds, it is challenging to determine a 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 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* tests. This design choice is justified, as individual tests are solvable by non-AGI systems, whereas the simultaneous satisfaction of all fourteen requirements is believed 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 axioms taken together.

3.3. AGITB axioms

Let $\mathbf{x} \in X = \{0,1\}^\omega$ with $\omega = 10$ denote a ten-bit input vector, and let $\boldsymbol{\chi} = (\mathbf{x}_1, \ldots, \mathbf{x}_{|\boldsymbol{\chi}|})$ denote a finite input sequence, where $|\boldsymbol{\chi}|$ is its length. A model instance under consideration is denoted by A, and a second, independent instance of the same model type by B. We write A_t for A at timestep t. For brevity, we refer to a model instance simply as a "model" whenever no ambiguity arises.

A model producing a prediction for the next input is written as

$$A_t \rightsquigarrow \mathbf{x}^*_{t+1}.$$

Upon receiving the actual input $\mathbf{x}_{t+1} \in X$, the model updates according to

$$A_t \xrightarrow{\mathbf{x}_{t+1}} A_{t+1}.$$

For any model A and any input sequence $\boldsymbol{\chi}$, we denote by

$$A \xrightarrow{\boldsymbol{\chi}} A^{\boldsymbol{\chi}}$$

the resulting model obtained after processing all inputs in $\boldsymbol{\chi}$. If a model is iteratively fed its own predictions, it generates a stream of predicted inputs:

$$A_t \Rrightarrow (\mathbf{x}^*_{t+1}, \mathbf{x}^*_{t+2}, \ldots).$$

A model may be exposed to multiple input sequences in succession, or repeatedly to the same sequence:

$$A \xrightarrow{\boldsymbol{\chi_1}} A^{\boldsymbol{\chi_1}} \xrightarrow{\boldsymbol{\chi_2}} A^{\boldsymbol{\chi_1}\boldsymbol{\chi_2}}, \qquad A \xrightarrow{\boldsymbol{\chi}\boldsymbol{\chi}} A^{\boldsymbol{\chi}^2}.$$

In AGITB, "learning" refers to a model's ability to update itself in response to each incoming input so as to generate a correct prediction of the next one. In this behavioural sense, a model can learn to reproduce an entire sequence. We write $A \rhd \boldsymbol{\phi}$ to denote a model that can learn a sequence $\boldsymbol{\phi}$:

$$\forall\, A: \qquad A \rhd \boldsymbol{\phi} \;:=\; \exists\, n \in \mathbb{N}:\; A^{\boldsymbol{\phi}^n} \Rrightarrow \boldsymbol{\phi}.$$

At a minimum, the model under evaluation must implement a state transition $A_t \to A_{t+1}$ that captures the internal updates induced by processing the input $\mathbf{x}_{t+1}$, and must also provide a mechanism for predicting the subsequent input. In the reference implementation, these two functionalities are exposed via distinct API functions: one function ingests and integrates the input $\mathbf{x}_t$ while returning the model's prediction $\mathbf{x}^*_{t+1}$, whereas the other function solely retrieves the current prediction without modifying the internal state.

AGITB rests on a single structural assumption: a model has an internal state that fully determines its behaviour. The form of this state, its representation, and the rules governing its update are left completely unspecified. Introducing a more detailed formal specification, such as defining the form of the state or the update rule, would risk privileging particular implementations and would run counter to AGITB's aim of remaining as architecture-agnostic as possible.

While A_{t+1} depends on the prior internal state A_t, the effect of an input may or may not require multiple updates to propagate, depending on the internal architecture of the model under consideration. For example, in an

artificial neural network architecture, the state update $A_t \xrightarrow{x_{t+1}} A_{t+1}$ constitutes the first step in the inference process leading to the output x^*_{t+1+d}, where d denotes the depth of the network[2]; this many forward passes are required for the influence of input x_{t+1} to reach the output. For all $i < d$, the intermediate outputs x^*_{t+1+i} are therefore independent of x_{t+1}.

Verifying a model against AGITB requires comparing independently instantiated copies of the model under controlled scenarios. In this context, two model instances are considered identical if and only if they are identical in state; otherwise, they are distinct. Because the models are treated as black boxes, AGITB cannot assess this directly. A naïve alternative would compare observable predictions over an unbounded horizon or until a divergence occurs, but this approach is generally infeasible. Instead, AGITB relies on the model developer to provide an efficient and accurate mechanism for determining whether two model instances are identical or distinct.

Several AGITB requirements cannot be validated through a single externally administered test. To accommodate this limitation, the AGITB test suite is designed so that individual tests act synergistically: while no single test can confirm a requirement in isolation, the collective behaviour observed across multiple tests enables indirect validation. Conformance with each of the following axioms can therefore be assessed by tests that verify the relevant properties of the model type under assessment. A deterministic C++ implementation serves as the benchmark's reference code and is freely available under the GPL-3 license at `https://github.com/matejsprogar/agitb`.

Axiom 1 (Bias-free start).
(a) All models begin in a completely blank, bias-free state:

$$\forall\, A :\ A_0 = \diamond.$$

(b) Blank models are identical:

$$\forall\, A, B :\ A_0 = B_0.$$

(c) An unbiased model produces an empty initial prediction:

$$\forall\, A :\ A_0 \rightsquigarrow \mathbf{0}.$$

[2]Here, depth is defined as the number of hidden layers plus the output layer.

An *unbiased* model A_0 is defined as an initial model with no inductive biases: it has no prior information about what it will encounter, no task-specific architectural priors, and no encoded expectations; it only has the structural capacity to learn.

To ensure neutrality, models must initialise in a completely blank state and remain so until conditioned on actual data. This bias-free start complements other requirements that operationalise how models acquire meaning in the absence of prior grounding.

The first clause is not externally verifiable and cannot be confirmed through a single test instance. Nevertheless, the testbed guarantees that no universal task-relevant prior knowledge exists, so embedding any such prior information would provide no advantage to any model.

Two blank model instances are considered equal, since neither has been shaped by prior experience. Additional expectations of equality or inequality are formalised in later axioms.

The final clause states that all bias-free models produce the same ω-bit zero vector $\mathbf{0}$ as their initial prediction, independent of the first input. A model that produces a non-empty prediction prior to receiving input is considered biased because it encodes assumptions about the future. Although an all-zero prediction could be construed as a bias toward silence, the convention is justified by biological analogy: neurons remain inactive in the absence of stimulation.

AGITB assumes that general-purpose learning systems, including biological brains, 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 knowledge.[3]

Axiom 2 (Bias). *A change in state indicates bias.*

$$\forall\, A :\ A_1 \neq A_0$$

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

[3]Although certain reflexes may be genetically specified, they do not constitute genuine understanding. Such reflexes are evolutionary features of the subcortical "old brain" and not prerequisites for intelligence [12, p. 66].

Axiom 3 (Determinism). *Identical inputs guarantee identical models.*

$$\forall\, \boldsymbol{\chi} :\ A = B \ \Rightarrow\ A^{\boldsymbol{\chi}} = B^{\boldsymbol{\chi}}.$$

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 identical model instances must have experienced identical input histories: any difference in input necessarily produces a divergence in state. Thus, the model 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 [13].

Axiom 4 (Sensitivity). *Distinct models remain distinct under identical inputs.*

$$\forall\, \boldsymbol{\chi} :\ A \neq B \ \Rightarrow\ A^{\boldsymbol{\chi}} \neq B^{\boldsymbol{\chi}}.$$

Sensitivity can be characterised without introducing a distance metric between models. Distinct models remain distinct regardless of the amount of identical input; they never converge to the same model.

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.

Corollary 1 (Injective determinism).

$$\forall\, \boldsymbol{\chi} :\ A = B \ \Leftrightarrow\ A^{\boldsymbol{\chi}} = B^{\boldsymbol{\chi}}.$$

Proof. The forward implication follows from Axiom 3. For the converse, assume $A^{\boldsymbol{\chi}} = B^{\boldsymbol{\chi}}$. By the contrapositive of Axiom 4, this implies $A = B$.

Axiom 5 (Time). *System evolution depends on input order.*

$$\forall\, \boldsymbol{\chi}_1 \neq\ \boldsymbol{\chi}_2 \in X^+ \ :\ A^{\boldsymbol{\chi}_1\boldsymbol{\chi}_2} \neq A^{\boldsymbol{\chi}_2\boldsymbol{\chi}_1}.$$

Because the model depends on its cumulative history, the sequence in which inputs are presented critically shapes its learning and adaptation. The ability to recognise and exploit temporal structure is regarded as a defining property of intelligent systems.

Axiom 6 (Absolute refractory period). *A model can learn a cyclic sequence only if the sequence satisfies the absolute refractory-period constraint.*

$$A \rhd \boldsymbol{\chi} \;\Rightarrow\; \boldsymbol{\chi} \in \boldsymbol{\Xi}.$$

The set of *admissible* sequences is defined as

$$\boldsymbol{\Xi} = \Big\{ \boldsymbol{\chi} = (\mathbf{x}_1, \ldots, \mathbf{x}_k) \in X^k,\; k < k_{max} \;:\; \forall i,\; \mathbf{x}_i \wedge \mathbf{x}_{((i \bmod k)+1)} = \mathbf{0} \Big\}.$$

Conversely, *learnable sequences* are a subset $\boldsymbol{\Phi} \subset \boldsymbol{\Xi}$ such that

$$\boldsymbol{\phi} \in \boldsymbol{\Phi} \;\Rightarrow\; A \rhd \boldsymbol{\phi}.$$

Biological intelligence relies on discrete spiking events for communication and learning, and individual neurons cannot fire again immediately after activation. AGITB incorporates an absolute refractory-period constraint to reflect this property, without assuming that all sequences admissible under this constraint are necessarily learnable.

Although absolute refractory periods are not themselves the source of spiking variability, they impose a minimum separation between spikes and thereby preserve temporal structure. Learning mechanisms based on spike timing cannot operate effectively when such a structure is absent [14]. Consequently, AGITB admits all temporal sequences consistent with biologically plausible refractory dynamics, while remaining agnostic to any particular semantic encoding of signals.

This axiom requires cyclic sequences, the simplest learning setting due to repeated input exposure; learning from non-repetitive input streams is addressed by Axiom 13.

Axiom 7 (Temporal adaptability). *The model must be able to learn sequences with varying cycle lengths.*

$$\exists\, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \;:\; 0 < |\boldsymbol{\phi}_1| < |\boldsymbol{\phi}_2| \;\wedge\; A \rhd \boldsymbol{\phi}_1 \;\wedge\; A^{\boldsymbol{\phi}_1} \rhd \boldsymbol{\phi}_2.$$

This axiom 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.

Axiom 8 (Bounded learnability).
(a) The model has limited learnability.

$$\forall A,\ \forall \boldsymbol{\phi}_0 \subset \boldsymbol{\Phi},\ \exists \boldsymbol{\phi}_1 \subset \boldsymbol{\Phi} :\ \neg\big(A^{\boldsymbol{\phi}_1} \rhd \boldsymbol{\phi}_0\big).$$

(b) All admissible length-2 sequences are universally learnable.

$$\forall A,\ \forall \boldsymbol{\chi} \in \boldsymbol{\Xi}\ :\ |\boldsymbol{\chi}| = 2\ \Rightarrow\ A \rhd \boldsymbol{\chi}.$$

Learning systems with finite representational and adaptive capacity exhibit inherent limits on the sets of sequences they can learn. In particular, no model can be trained so as to learn all learnable sequences. However, a nontrivial lower bound is preserved: every model can learn any learnable sequence of length two.

Axiom 9 (Content sensitivity). *Adaptation time depends on the content of the input sequence.*

$$\exists\, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2 :\quad |\boldsymbol{\phi}_1| = |\boldsymbol{\phi}_2|\ \wedge\ t_{A \rhd \boldsymbol{\phi}_1} \neq t_{A \rhd \boldsymbol{\phi}_2}.$$

where $t_{A \rhd \boldsymbol{\phi}}$ denotes the number of timesteps required for the model to learn the sequence $\boldsymbol{\phi}$.

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.

Axiom 10 (Context sensitivity). *Adaptation time depends on the state of the model.*

$$\exists\, \boldsymbol{\phi},\ A,\ B :\ A \neq B\ \wedge\ t_{A \rhd \boldsymbol{\phi}}\ \neq\ t_{B \rhd \boldsymbol{\phi}}.$$

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.

Axiom 11 (Unobservability). *Inequivalent models may exhibit the same observable behaviour.*

$$\exists\, \boldsymbol{\phi},\, A,\, B:\ A \neq B\ \wedge\ A \rhd \boldsymbol{\phi}\ \wedge\ B \rhd \boldsymbol{\phi}.$$

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.

Axiom 12 (Denoising). *The model performs above chance on perturbed inputs.*

$$\forall\, \boldsymbol{\phi}:\ A \rhd \boldsymbol{\phi}\ \wedge\ A^{\boldsymbol{\phi}'} \rightsquigarrow \mathbf{x}_1^*\ \Rightarrow\ P(\mathbf{x}_1^*[i] = \mathbf{x}_1[i]) > \tfrac{1}{2},$$

where $\boldsymbol{\phi}' = (\mathbf{x}_1', \mathbf{x}_2, \ldots, \mathbf{x}_k)$ *is a perturbed version of* $\boldsymbol{\phi}$ *with* $\mathbf{x}_1' \neq \mathbf{x}_1$.

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.

The corresponding test is intentionally designed to be conservative: the model must outperform random guessing in each of 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.

Axiom 13 (Generalisation). *The model performs above chance on previously unseen inputs.*

$$\forall A: \quad A \Rrightarrow (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) \;\Rightarrow$$
$$\Big(A_0^{\boldsymbol{\phi}_1} \Rrightarrow \boldsymbol{\phi}_2' \;\wedge\; P(\boldsymbol{\phi}_2'[i][j] = \boldsymbol{\phi}_2[i][j]) > \tfrac{1}{2}\Big).$$

Only models capable of generalisation can derive true learning benefits. 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.

Axiom 14 (Bounded state update).

$$\forall A, \exists \tau \in N \;: \quad t_{A_i \to A_{i+1}} < \tau.$$

A model must update its internal state in response to a single input within bounded time to remain functionally viable. In biological brains, a single cortical state update completes in approximately constant time due to massive parallelism, keeping per-input state update time bounded and stable across expected operating conditions.

To verify this requirement in a model, the corresponding test compares state update times between two configurations: an empty model and a complex model, evaluated under identical input conditions. If the bounded state update axiom holds, the complex model should not exhibit state update times that are consistently greater than those of the empty model. This condition is assessed using a one-sided Wilcoxon signed-rank test with a very conservative threshold of $z = 3.090$, corresponding to a one-sided significance level of 0.1%. This stringent setting strongly limits false positives and ensures reliable detection of models that violate the axiom.

3.4. Search space

To prevent models from relying on brute-force memorisation, a robust 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 require predicting a temporal sequence of seven inputs ($|\boldsymbol{\phi}| = 7$), each consisting of ten bits ($\omega = 10$). This corresponds to a combinatorial space of size $|S| = 2^{70}$, representing all possible binary input sequences of that length.

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 $|S'| = (F_{|\boldsymbol{\phi}|+2})^{\omega} = 34^{10} \approx 2^{51}$ 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 $|S''| = (L_{|\boldsymbol{\phi}|})^{\omega} = 29^{10} \approx 2^{49}$, 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. Interpreting performance under AGITB

Before assessing the usefulness of AGITB, it is important to clarify its role as a pragmatic benchmark rather than an end in itself. Like the Turing Test, which serves as an empirical criterion rather than a philosophical claim

[15], AGITB is intended as a practical instrument for evaluating progress toward artificial general intelligence. The ultimate objective remains the development of AGI, not merely success on the benchmark.

AGITB 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, can 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, it is necessary to establish a baseline by considering the performance of human cognition.

3.5.1. Human performance

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

Owing to prior experience and cognitive bias, an adult human's cortex may appear to "fail" this 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 (Axiom 7), for example, poses no difficulty for humans, who readily recognise temporal patterns of varying durations. Bounded learnability (Axiom 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 (Axiom 9) and context sensitivity (Axiom 10) is immediately evident. The unobservability test (Axiom 11) formalises the fact that internal human

mental states cannot be directly accessed or inferred from behaviour alone.

The denoising test (Axiom 12) and the generalisation test (Axiom 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 (Axiom 14) is a well-established property of the human brain, which maintains approximately constant reaction times through extensive parallel processing.

3.5.2. Classical symbolic program performance

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

The former category encompasses most AI and purported AGI systems developed to date; however, it remains fundamentally constrained by the Symbol Grounding Problem (SGP) [8]. 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, biased systems 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 formalises the requirement to learn from scratch, demanding that a system derive structure and meaning solely through exposure to intrinsically grounded binary signals. Only a bias-free system could, in principle, satisfy this condition. This suggests that a genuine AGI may not explicitly encode the operations of intelligence, but rather the dynamics of a substrate from which they can emerge. This aligns with the idea of the "Brain Simulator Reply" [16] to Searle's Chinese Room Argument [17]. To date, however, no such system has been demonstrated.

3.5.3. Artificial neural network performance

The expectation dichotomy also extends to connectionist architectures. In principle, artificial neural networks may incorporate built-in expectations, introduced through pretraining regimes or architectural priors, or they may be configured as expectation-free systems that start from a neutral initial state.

AGITB's requirement that models begin in a bias-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 such 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. Internal model coherence does not entail external semantic validity.

An expectation-free network carries no inductive bias in the form of weights or architectural priors. Yet a standard neural network without initialised weights cannot engage in effective learning, as its units remain functionally dormant and generate no informative output. This highlights a fundamental limitation of current ANN architectures: they do not learn autonomously but instead depend on an external training procedure to drive adaptation. AGITB, by contrast, requires a blank system capable of autonomous adaptation in an unfamiliar environment. To date, no such mechanism has been demonstrated in artificial neural networks.

3.5.4. Large language model performance

Because a large language model (LLM) is a form of artificial neural network, it fails the bias-free start test (Axiom 1). LLMs cannot learn entirely from scratch, as their behaviour is predetermined by the linguistic patterns and data used during pre-training. Although an LLM's internal state is fully accessible and, unlike that of a human, can be examined or compared across instances, such evaluations are moot until the first requirement is satisfied.

A further limitation arises from the LLM architecture itself, which relies on a fixed-size context window [18]. Once its capacity is exceeded, earlier tokens are discarded, compressed, or attenuated [19]. This violates the sen-

sitivity requirement (Axiom 3), which stipulates that a model must remain responsive to any individual input even after processing an arbitrarily long sequence of subsequent inputs.

Nevertheless, the question of whether an LLM can autonomously derive a solution to AGITB when prompted is straightforward to evaluate empirically. Multiple attempts using distinct prompts (one example is provided in Appendix A) yielded unpromising results: although systems such as ChatGPT, Gemini or Claude produced candidate programs that purported to satisfy the requirements, none progressed beyond the temporal adaptability test (Axiom 7).

In summary, LLMs do not engage in 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 inherit the same fundamental architectural constraints.

3.6. 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 do not need to resemble real-world sensory data.

The low-level, binary operational framework makes AGITB particularly well-suited for evaluating NeuroAI models that aim to satisfy the principles of the embodied Turing Test [20], 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.

3.7. 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, thereby hindering 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` in the reference implementation, 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 [5] 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 alleviate the broader symbol-grounding problem in general.

In summary, ARC evaluates high-level intelligence grounded in human cognitive priors, whereas AGITB evaluates adherence to fourteen low-level

computational requirements 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 [21].

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.

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
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 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 learning expected of an AGI.

AGITB requires models to begin in a bias-free 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.

Funding

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

Appendix A. Model construction prompt

You are an expert engineer and sequence-learning researcher. Your task is to create a **concrete solution** that satisfies the AGITB benchmark defined in the accompanying .zip archive.

Produce complete, compilable C++20 source code, including all required classes, methods, and internal logic needed to satisfy the benchmark's tests.

1. Study the following files:
– `README.md`
– `include/agitb.h`
– `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 instructions above to guide your analysis and implementation.

References

[1] G. Marcus, Deep learning: A critical appraisal (2018). arXiv:1801.00631.
URL `https://arxiv.org/abs/1801.00631`

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

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

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

[5] F. Chollet, On the measure of intelligence (2019). arXiv:1911.01547.
URL `https://arxiv.org/abs/1911.01547`

[6] J. Yik, K. V. den Berghe, D. den Blanken, Y. Bouhadjar, M. Fabre, P. Hueber, W. Ke, M. A. Khoei, D. Kleyko, N. Pacik-Nelson, A. Pierro, P. Stratmann, P.-S. V. Sun, G. Tang, S. Wang, B. Zhou, S. H. Ahmed, G. V. Joseph, B. Leto, A. Micheli, A. K. Mishra, G. Lenz, T. Sun, Z. Ahmed, M. Akl, B. Anderson, A. G. Andreou, C. Bartolozzi, A. Basu, P. Bogdan, S. Bohte, S. Buckley, G. Cauwenberghs, E. Chicca, F. Corradi, G. de Croon, A. Danielescu, A. Daram, M. Davies, Y. Demirag,

J. Eshraghian, T. Fischer, J. Forest, V. Fra, S. Furber, P. M. Furlong, W. Gilpin, A. Gilra, H. A. Gonzalez, G. Indiveri, S. Joshi, V. Karia, L. Khacef, J. C. Knight, L. Kriener, R. Kubendran, D. Kudithipudi, S.-C. Liu, Y.-H. Liu, H. Ma, R. Manohar, J. M. Margarit-Taulé, C. Mayr, K. Michmizos, D. R. Muir, E. Neftci, T. Nowotny, F. Ottati, A. Ozcelikkale, P. Panda, J. Park, M. Payvand, C. Pehle, M. A. Petrovici, C. Posch, A. Renner, Y. Sandamirskaya, C. J. S. Schaefer, A. van Schaik, J. Schemmel, S. Schmidgall, C. Schuman, J. sun Seo, S. Sheik, S. B. Shrestha, M. Sifalakis, A. Sironi, K. Stewart, M. Stewart, T. C. Stewart, J. Timcheck, N. Tömen, G. Urgese, M. Verhelst, C. M. Vineyard, B. Vogginger, A. Yousefzadeh, F. T. Zohora, C. Frenkel, V. J. Reddi, The neurobench framework for benchmarking neuromorphic computing algorithms and systems, Nature Communications 16 (1) (2025) 1545. doi:10.1038/s41467-025-56739-4.

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

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

[9] E. M. Bender, A. Koller, Climbing towards NLU: On meaning, form, and understanding in the age of data, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5185–5198. doi:10.18653/v1/2020.acl-main.463.
URL https://aclanthology.org/2020.acl-main.463/

[10] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich, N. Pinto, J. Turian, Experience grounds language, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 8718–8735. doi:10.18653/v1/2020.emnlp-main.703.
URL https://aclanthology.org/2020.emnlp-main.703/

[11] R. Gubelmann, Pragmatic norms are all you need – why the symbol grounding problem does not apply to LLMs, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 11663–11678. doi:10.18653/v1/2024.emnlp-main.651.
URL https://aclanthology.org/2024.emnlp-main.651/

[12] J. Hawkins, S. Blakeslee, On Intelligence, Times Books, 2004.

[13] 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/

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

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

[16] P. Churchland, P. Churchland, Could a machine think?, Scientific American 262 (1) (1990) 32–37.

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

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.
URL http://arxiv.org/abs/1706.03762

[19] N. Paulsen, Context is what you need: The maximum effective context window for real world limits of llms (2025). arXiv:2509.21361.
URL https://arxiv.org/abs/2509.21361

[20] 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, D. Tsao, Catalyzing next-generation artificial intelligence through NeuroAI, Nature Communications 14 (1) (2023) 1597. doi:10.1038/s41467-023-37180-x.

[21] M. Thill, W. Konen, T. Bäck, Time series encodings with temporal convolutional networks, in: B. Filipič, E. Minisci, M. Vasile (Eds.), Bioinspired Optimization Methods and Their Applications, Springer International Publishing, Cham, 2020, pp. 161–173. doi:10.1007/978-3-030-63710-1_13.