

Intelligence per Watt: A Brain-Inspired Framework for Continual On-Device Learning on Low-Cost Hardware

Oct 13, 2026 · @Munish Dhiman

Munish Dhiman, 86b.ai, India

Article type: Perspective

Abstract

The human brain supports general intelligence with roughly 86 billion neurons on about 20 watts, while state-of-the-art artificial systems require orders of magnitude more energy and stop learning once deployed. We argue that the next gains in practical AI will come not only from scale but from intelligence per watt: systems that learn continually, compute sparsely and adapt to individual users on the devices they already own. We propose a framework that combines four brain-inspired mechanisms: complementary fast and slow learning systems, sparse conditional computation, predictive self-supervised learning and uncertainty-gated adaptation. The framework is designed for on-device continual learning on low-cost smartphones. We specify an evaluation protocol that measures accuracy, forgetting, energy per inference and per update, latency, memory and data efficiency on real budget hardware rather than in simulation. We further propose an Indian-context benchmark spanning multiple languages, accents, acoustic and lighting conditions and low-cost device tiers, arguing that conditions typical of India represent those of most of the world's users and are under-represented in current evaluations. We state falsifiable hypotheses, outline a personalised voice-command task as the first testbed, and discuss limitations and ethical safeguards. No experimental results are reported here; this paper defines the problem, the approach and the evaluation standard against which results will be reported.

Keywords: continual learning, catastrophic forgetting, on-device learning, energy-efficient AI, brain-inspired computing, sparse computation, edge AI, low-resource settings

1. Introduction

Progress in artificial intelligence over the past decade has been driven largely by scale: larger models, larger datasets and larger compute budgets. This strategy has produced remarkable capabilities, but it carries three costs. First, training and serving frontier models demands energy and capital available to only a few organisations. Second, the resulting models are poorly suited to the phones, wearables and embedded devices where most people interact with technology. Third, deployed models are largely static: they do not continue to learn from the users and environments they encounter, and naive attempts to update them cause catastrophic forgetting of earlier knowledge [McCloskey and Cohen, 1989].

The brain offers an existence proof that a different operating point is possible. The human brain contains approximately 86 billion neurons [Herculano-Houzel, 2009] and operates on roughly 20 watts [Attwell and Laughlin, 2001]. It learns continually from limited experience, activates only a small fraction of its neurons at any time, and retains old skills while acquiring new ones. These properties are precisely those that on-device AI lacks.

This gap matters most for the majority of the world's users. In India and comparable regions, most people use low-cost smartphones, often shared, in variable acoustic and lighting conditions, across many languages and dialects, with intermittent connectivity. Models trained and evaluated on flagship hardware and narrow populations frequently degrade in these conditions, and cloud-based personalisation raises cost, latency and privacy concerns.

We argue for intelligence per watt as a first-class research objective, and make the following contributions:

1. A framework combining four brain-inspired mechanisms for continual on-device learning on low-cost hardware (Section 3).
2. An evaluation protocol that reports accuracy jointly with forgetting, energy, latency, memory and data efficiency, measured on real budget devices (Section 4).
3. A proposed Indian-context benchmark capturing linguistic, acoustic, environmental and device diversity typical of most of the world's users (Section 5).
4. Falsifiable hypotheses against which the framework will be tested (Section 6).

2. Background and related work

Continual learning. Neural networks trained sequentially on new data tend to overwrite earlier knowledge, a phenomenon termed catastrophic interference [McCloskey and Cohen, 1989]. Approaches fall into three families [Parisi et al., 2019]. Regularisation methods protect parameters important to earlier tasks, as in elastic weight consolidation [Kirkpatrick et al., 2017] and synaptic intelligence [Zenke et al., 2017]. Replay methods rehearse stored or generated examples, as in iCaRL [Rebuffi et al., 2017] and gradient episodic memory [Lopez-Paz and Ranzato, 2017]. Architectural methods allocate new capacity for new knowledge. Most of this work is evaluated on servers with generous memory and compute budgets, not on the devices where continual learning would be most useful.

Complementary learning systems. Neuroscience suggests the brain avoids catastrophic forgetting by pairing a fast-learning hippocampal system with a slow-learning neocortical system, with memories consolidated gradually through replay [McClelland et al., 1995]. This division of labour directly inspires our fast and slow learning components.

Sparse and conditional computation. Only a small fraction of cortical neurons is active at any moment, which is central to the brain’s energy efficiency [Attwell and Laughlin, 2001]. In AI, mixture-of-experts models activate a subset of parameters per input [Shazeer et al., 2017], and model compression reduces cost through pruning and quantisation [Han et al., 2016]. These techniques have mostly targeted large-scale inference rather than small-device learning.

Predictive coding. The theory that the cortex continually predicts its inputs and learns from prediction errors [Rao and Ballard, 1999] motivates self-supervised objectives that extract structure from unlabelled data, reducing dependence on labels during adaptation.

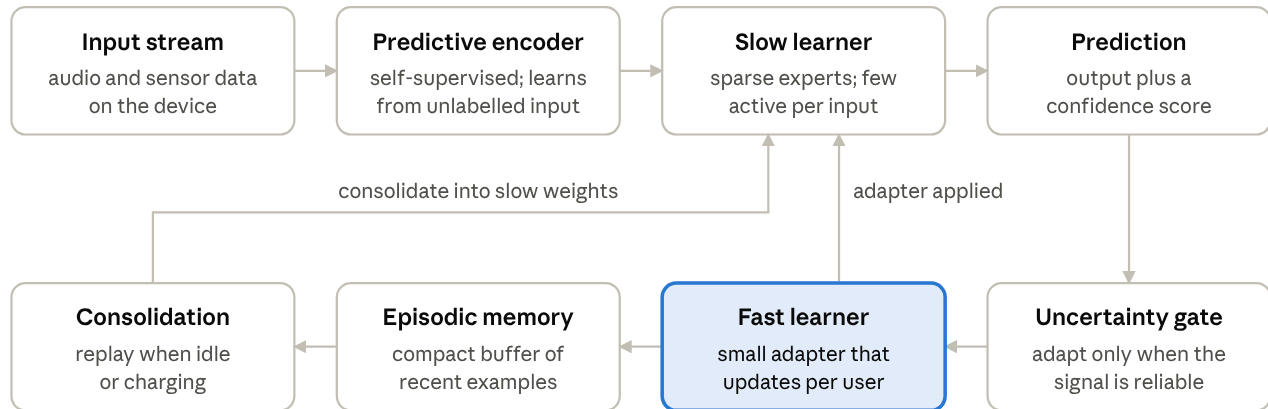
Spiking neural networks. Spiking models communicate through discrete events, as biological neurons do [Maass, 1997], and promise large energy savings on neuromorphic and event-driven hardware. Their accuracy and tooling still lag conventional networks on many tasks.

On-device learning. Recent work has shown that training is feasible under very tight memory budgets on microcontrollers [Lin et al., 2022]. However, the combination of continual learning, energy measurement on real low-cost smartphones, and evaluation across diverse populations remains largely unexplored. This is the gap our framework addresses.

3. The proposed framework

The framework separates what a model knows from how it adapts, so that each user's device can learn quickly without overwriting shared knowledge.

Fast per-user learning is consolidated into a sparse slow learner



proposed framework · slow path (top), fast path (bottom)

The top row is the inference path; the bottom row is the adaptation loop, which updates only a small fast learner and consolidates into the slow learner when the device is idle.

3.1 Complementary fast and slow learners. Following complementary learning systems theory [McClelland et al., 1995], the model has two parts. The slow learner holds general knowledge and changes rarely. The fast learner is a small, low-rank adapter that updates online from each user's data. Because day-to-day learning touches only the adapter, the shared knowledge in the slow learner is protected from catastrophic forgetting.

3.2 Sparse conditional computation. The slow learner is organised as a set of compact experts with a lightweight router that activates only a few experts per input [Shazeer et al., 2017]. This reduces computation per inference and limits which parameters each update can affect, a further guard against interference.

3.3 Predictive self-supervised encoding. Inspired by predictive coding [Rao and Ballard, 1999], the encoder is trained to predict masked or future segments of its input. This lets the device learn useful structure from unlabelled audio and sensor streams, reducing the number of labelled user examples adaptation needs.

3.4 Uncertainty-gated adaptation. Each prediction carries a calibrated confidence score. The fast learner updates only when the signal is reliable, for example when a user confirms or corrects a prediction, or when confidence is high and consistent over time.

This prevents the model from learning its own mistakes or from noisy, shared-device inputs.

3.5 Episodic memory and consolidation. A compact buffer stores a small number of recent examples. When the device is idle or charging, a consolidation step replays these examples together with stored summaries of earlier knowledge to transfer stable improvements from the fast learner into the slow learner, analogous to memory consolidation during rest. The buffer is bounded and stays on the device.

3.6 Optional event-driven execution. For always-on sensing, the encoder can be implemented as a spiking or event-driven network [Maass, 1997] that computes only when input changes, for deployment on neuromorphic or low-power hardware. This is an extension, not a requirement of the core framework.

4. Evaluation protocol

We propose that on-device continual learning be evaluated on real low-cost hardware, with accuracy always reported alongside its costs.

Tasks. The first testbed is personalised voice-command recognition: a compact keyword and command model, pre-trained on a public corpus such as Speech Commands [Warden, 2018], then adapted on the device to individual speakers across languages and accents. Standard continual-learning benchmarks such as CORE50 [Lomonaco and Maltoni, 2017] provide comparability with published methods. Sensor-based activity recognition serves as a second, non-audio task.

Baselines. Each method is compared at equal parameter count and memory budget with: (a) a static pre-trained model with no adaptation; (b) naive fine-tuning; (c) elastic weight consolidation [Kirkpatrick et al., 2017]; (d) experience replay with a fixed buffer; and (e) a compressed dense model of the same size [Han et al., 2016].

Metrics.

Metric	Definition	Unit
Adapted accuracy	Accuracy on the individual user's held-out data after adaptation	%
Forgetting	Average drop in accuracy on earlier tasks or users (backward transfer [Lopez-Paz and Ranzato, 2017])	percentage points

Metric	Definition	Unit
Energy per inference	Measured device energy for one prediction	mJ
Energy per update	Measured device energy for one adaptation step	mJ
Latency	Time per inference and per update on the device	ms
Peak memory	Maximum RAM used during inference and adaptation	MB
Data efficiency	User examples needed to reach a target accuracy	count
Calibration	Expected calibration error of confidence estimates	%

Devices. Experiments run on at least three Android smartphones in the low-cost tier, with different chipset families, alongside one single-board computer. Energy is measured with an external power monitor rather than estimated from software counters, and each figure is reported as a mean with confidence interval over repeated runs.

Reporting. Results are reported per device and per user subgroup, not only as averages, so that degradation for specific populations or hardware is visible.

5. Evaluating for the world's majority: an Indian-context benchmark

We propose a benchmark built in India because conditions typical of Indian users are the conditions of most of the world's users, and they are rarely represented in current evaluations.

Most on-device AI is developed and validated on flagship hardware, in quiet or well-lit conditions, with speakers of a few high-resource languages. A method that performs well there may fail where it is most needed. India offers a natural test bed: more than 20 major languages and many more dialects, widespread code-mixed speech, a large share of users on low-cost devices, shared phones, and highly variable acoustic and lighting environments.

Proposed benchmark dimensions

Dimension	Variation captured
Language	Multiple Indian languages, including lower-resource ones, plus code-mixed speech
Speaker	Regional accents, age groups, gender balance
Environment	Quiet rooms, street noise, households, markets, vehicles
Device tier	Low-cost, mid-range and reference flagship phones across chipset families
Usage	Single-user and shared-device scenarios
Connectivity	Fully offline operation as the default condition

Collection principles. Data is collected with informed consent, under approval from an institutional ethics committee, with fair compensation for participants and with speakers' identities protected. Released subsets are licensed for research use, and the composition of each subset is documented so that results can be interpreted by subgroup.

The benchmark is intended as a contribution in its own right: an evaluation standard that asks whether efficient, adaptive AI works for the populations that make up the global majority, developed in the region where those conditions are the norm.

6. Hypotheses and expected contributions

We state the hypotheses in falsifiable form; each will be accepted or rejected against the baselines and metrics of Section 4.

- **H1 (forgetting).** At equal parameter count and memory budget, the fast and slow learning design shows lower forgetting than naive fine-tuning and elastic weight consolidation, and matches or improves on replay with a fixed buffer.
- **H2 (energy).** Sparse conditional computation reduces measured energy per inference and per update on low-cost phones relative to a dense model of equal size, without a statistically significant loss in adapted accuracy.
- **H3 (data efficiency).** Predictive self-supervised pre-training reduces the number of labelled user examples needed to reach a target accuracy.

- **H4 (equity).** Personalised on-device adaptation narrows the accuracy gap between best- and worst-performing speaker subgroups, languages and device tiers compared with a static model.
- **H5 (reliability).** Uncertainty-gated updates reduce harmful adaptations, measured as fewer accuracy regressions after updates, compared with ungated adaptation.

If these hypotheses hold, the expected contributions are a reference architecture for continual learning on low-cost devices, an open evaluation protocol that makes energy and forgetting first-class metrics, and a benchmark for adaptive AI in conditions typical of the global majority. If they do not, the protocol and benchmark remain useful as a standard for evaluating alternative approaches.

7. Limitations and ethical considerations

Biological analogy. The framework is inspired by, not a model of, the brain. Mechanisms such as complementary learning systems are used as design principles; we make no claim that the implementation reproduces biological computation, and its value must be judged on the engineering metrics alone.

Scope. The proposed testbeds are narrow, compact-model tasks. Results may not transfer to large generative models or to tasks requiring broad world knowledge. Gains measured on specific phones and chipsets may vary across hardware generations.

Measurement. Device energy depends on background processes, thermal state and operating-system scheduling. We mitigate this with external power monitoring, controlled conditions and repeated runs, but absolute figures should be read as indicative and compared within the same setup.

Privacy. On-device learning keeps raw user data local by default, but adapted model parameters can still leak information about the user. Future work should assess membership-inference risk and consider differential privacy for any parameters that leave the device.

Fairness. Personalisation could improve performance for frequent users while leaving occasional or shared-device users behind. Subgroup reporting (Section 4) is designed to detect this.

Data ethics. Benchmark collection follows informed consent, ethics committee approval, fair compensation and data protection consistent with India's Digital Personal Data Protection Act, 2023.

8. Conclusion

The brain demonstrates that continual, general-purpose learning is possible on about 20 watts. Current AI is far from that operating point, and the gap is widest on the low-cost devices used by most of the world's population. We have proposed a framework that brings four brain-inspired mechanisms to on-device continual learning, an evaluation protocol that treats energy and forgetting as first-class metrics on real hardware, and a benchmark grounded in the linguistic, environmental and device diversity of India. We invite the community to test, challenge and extend these hypotheses, and we will report results against this protocol as experiments proceed.

Declaration of competing interests. The author is the founder of 86b.ai, a company developing technology related to this work.

Funding. This work received no specific grant from any funding agency.

Data availability. No data were generated or analysed for this perspective. The proposed benchmark will be released with documentation upon collection.

References

1. Attwell, D. and Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism*, 21(10).
2. Han, S., Mao, H. and Dally, W. J. (2016). Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. *ICLR*.
3. Herculano-Houzel, S. (2009). The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3.
4. Kirkpatrick, J. et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13).
5. Lin, J. et al. (2022). On-device training under 256KB memory. *NeurIPS*.
6. Lomonaco, V. and Maltoni, D. (2017). COrE50: a new dataset and benchmark for continuous object recognition. *Conference on Robot Learning (CoRL)*.
7. Lopez-Paz, D. and Ranzato, M. (2017). Gradient episodic memory for continual learning. *NeurIPS*.
8. Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural Networks*, 10(9).

9. McClelland, J. L., McNaughton, B. L. and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex. *Psychological Review*, 102(3).
10. McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. *Psychology of Learning and Motivation*, 24.
11. Parisi, G. I. et al. (2019). Continual lifelong learning with neural networks: a review. *Neural Networks*, 113.
12. Rao, R. P. N. and Ballard, D. H. (1999). Predictive coding in the visual cortex. *Nature Neuroscience*, 2(1).
13. Rebuffi, S.-A. et al. (2017). iCaRL: incremental classifier and representation learning. *CVPR*.
14. Shazeer, N. et al. (2017). Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. *ICLR*.
15. Warden, P. (2018). Speech Commands: a dataset for limited-vocabulary speech recognition. *arXiv:1804.03209*.
16. Zenke, F., Poole, B. and Ganguli, S. (2017). Continual learning through synaptic intelligence. *ICML*.