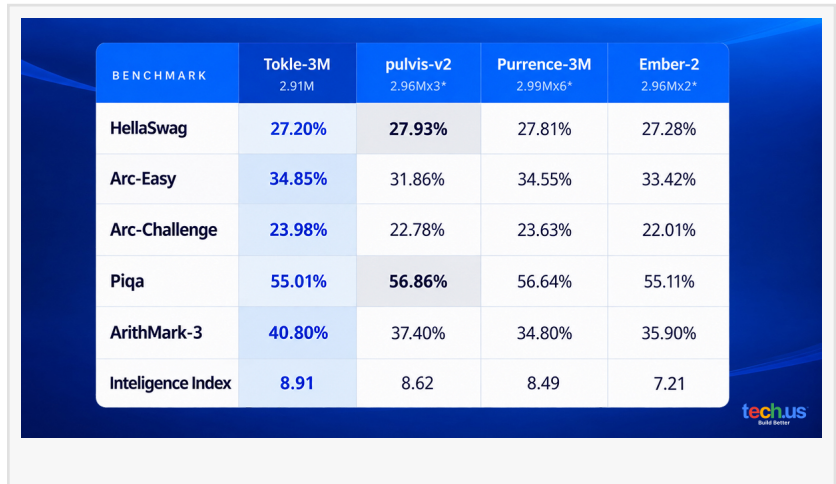


Tech.us' Tokle-3M Ranks #1 Among Sub-3M-Parameter Models on the Open SLM Leaderboard

SAN JOSE, CA, UNITED STATES, October 9, 2026 /EINPresswire.com/ -- The open model uses a guided-training technique to post the highest Intelligence Index score in its size class, with no added parameters at inference

[Tech.us](#) released today [Tokle-3M](#), an open small language model with 2.91 million parameters that ranks first among models under 3 million

parameters on the Open SLM Leaderboard. As of Oct. 1, 2026, Tokle-3M scores 8.91 on the leaderboard's Intelligence Index, and posts the highest arithmetic score in its class at 40.80% on ArithMark-3.



BENCHMARK	Tokle-3M 2.91M	pulvis-v2 2.96Mx3*	Purrence-3M 2.99Mx6*	Ember-2 2.96Mx2*
HellaSwag	27.20%	27.93%	27.81%	27.28%
Arc-Easy	34.85%	31.86%	34.55%	33.42%
Arc-Challenge	23.98%	22.78%	23.63%	22.01%
Piqa	55.01%	56.86%	56.64%	55.11%
ArithMark-3	40.80%	37.40%	34.80%	35.90%
Intelligence Index	8.91	8.62	8.49	7.21

“

We designed this model to be purpose-built from the start, with guided training helping it make better use of what it learns.”

Praveen Narra, CEO of Tech.us

Measured against the leaderboard's trend line for performance versus model size, Tokle-3M scores 1.80 standard deviations above what its parameter count would predict, the strongest result of any organization in the category. Tech.us attributes that gap to Static Pairwise Attention Bias (SPAB), a training approach that guides the model during learning and is then removed before release. Full training and evaluation details for Tokle-3M are published on the Tech.us research site.

The research focuses specifically on models in a comparable size range rather than positioning Tokle-3M as an alternative to large, general-purpose language models.

GIVING A SMALL MODEL A HEAD START

Tokle grew out of Tech.us' work on Static Pairwise Attention Bias, or SPAB, a training approach

designed to give a small model additional information about relationships between tokens.

During the first stage of training, SPAB uses statistical patterns in the training data to show the model which words or tokens are more likely to be related. That signal is then added to the model's attention process, helping it focus on useful relationships earlier in training instead of having to discover all of them on its own.

The second stage is where the approach becomes more distinctive. The SPAB information is removed, and what the model learned with that guidance is distilled into the final Tokle-3M model. The released model therefore runs with 2.91 million parameters without requiring the additional SPAB table at inference.

That gave Tech.us a way to test a simple research idea: whether a small model can be given richer guidance while it learns and retains much of the benefit once that guidance is taken away.

MEASURING WHAT A SMALL MODEL RETAINS

Tokle-3M was evaluated across HellaSwag, ARC-Easy, ARC-Challenge, PIQA, and ArithMark-3 using zero-shot normalized accuracy, and is ranked on the Open SLM Leaderboard. The leaderboard's Intelligence Index combines chance-adjusted results across reasoning, commonsense and arithmetic benchmarks into a composite measure.

The model was pre-trained on English datasets with a 512-token context window. Because it uses rotary position embeddings (RoPE), the window can be extended at inference time without changing the architecture, though performance beyond 512 tokens has not been evaluated. The model is not instruction-tuned or safety-aligned, and Tech.us is positioning the current release as a research model rather than a production assistant.

A DIFFERENT DIRECTION FOR MODEL EFFICIENCY

The research comes as the AI industry looks more closely at where model size is necessary and where it may not be.

Large general-purpose language models remain well suited to tasks that demand broad knowledge and flexible reasoning. But not every AI workload carries those requirements. A model designed around a narrow domain, defined workflow or constrained environment may have a very different balance of performance, memory, latency and infrastructure needs.

In its current form, Tokle-3M is intended for researchers and engineers studying how training methods affect compact models. Its size makes experiments fast and inexpensive to run, and its open weights let others reproduce and build on the results.

With task-specific fine-tuning, models in this class could support narrow, well-defined workloads such as text classification, intent detection, interpreting device commands and triaging system logs, including in offline, embedded or on-device settings where data must stay on local hardware.

Tech.us says its longer-term goal is models small enough that organizations can run them entirely on infrastructure they control. On-device and self-hosted deployment can help keep sensitive information within environments the organization already manages, reducing the need to move that data beyond those boundaries.

That is where Tech.us sees the longer-term relevance of its SLM research.

The company is exploring whether techniques developed through models such as Tokle can eventually contribute to more focused AI systems, particularly where organizations need tighter control over how models are trained, deployed, and adapted to specific work.

Tokle-3M provides an early technical result: in a deliberately compact architecture, the way a model is trained can make a measurable difference in performance. The company's broader aim is to understand how focused training, compact architecture and deployment control work together to make AI more practical for specific business environments.

AVAILABILITY

Tokle-3M is available now on [Hugging Face](#) along with a model card describing its training, evaluation, and limitations.

ABOUT Tech.us

Tech.us is an AI consulting and implementation company headquartered in San Jose, California. The company designs and builds production-ready AI, agentic AI and custom software systems for mid-market and enterprise organizations. Its AI research explores practical approaches to model efficiency, architecture and deployment alongside its work delivering AI systems for real-world business applications.

More information is available at [tech.us](#).

Alex Maria Moses M

Tech.us

+1 866-277-3383

ammoses@tech.us

Visit us on social media:

[LinkedIn](#)

[Instagram](#)

[Facebook](#)

This press release can be viewed online at: <https://www.einpresswire.com/article/948449976>

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.

© 1995-2026 Newsmatics Inc. All Right Reserved.