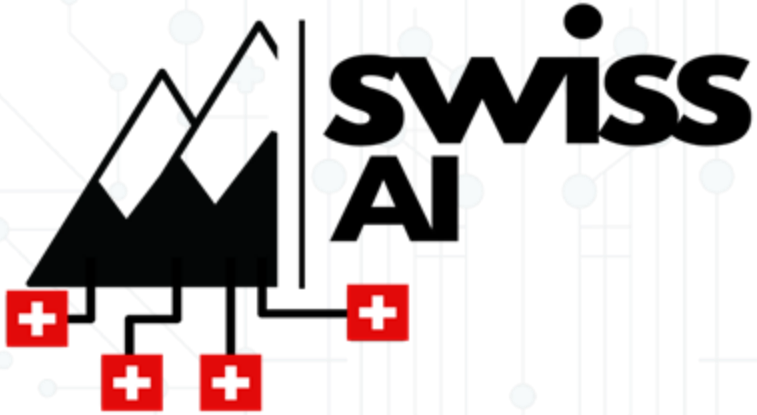


EPFL AI Center

ETH AI CENTER

Swiss AI writes many words at once

SwissAI-DLM

Faster, cheaper language AI, made in Switzerland

Toward a Swiss diffusion language model

Arbër Zela & Francesco Tonin

Swiss AI Industry Connect – 11.11.2026

PROJECT OVERVIEW

An LLM that writes many words at once

Faster answers
Diffusion LLMs already exceed 1,000 tokens/s.[1]

Lower cost per answer
More text per GPU-second, so each answer costs less.

Swiss, transparent, yours
Open data, trained in Switzerland, runs on-premise.

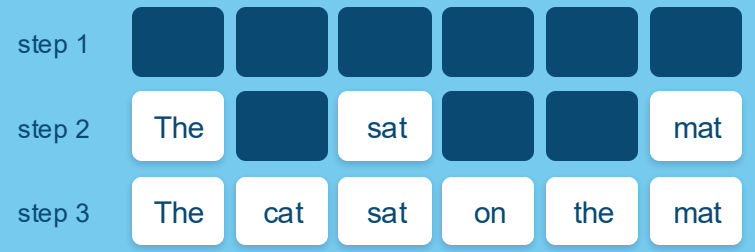
Backed by: Swiss AI Initiative, ~300K GPU-hours on Alps (CSCS)
Leads: Volkan Cevher (EPFL), Paolo Favaro (UniBE)
Ilijia Bogunovic, Aurelien Lucchi (UniBas), Andreas Krause (ETHZ)

[1] Vendor-reported figures: Google DeepMind, DiffusionGemma (Jun 2026); Inception Labs, Mercury 2 (Feb 2026).

Today's LLMs — one word per step



Our diffusion LLM — many per step



EPFL AI Center
ETH AI CENTER

SERVING GAINS

A diffusion LLM writes the same answer 2.7× faster

Question: Janet's ducks lay 16 eggs a day. She eats 3, bakes muffins with 4 and sells the rest for $2 each. How much does she make a day?

Standard LLM: one word per step

8B parameters · 229 tokens/s

Janet starts with 16 eggs a day. She eats 3 and bakes with 4, so 16 − 3 − 4 = 9 eggs are left. She sells them for $2 each: 9 × $2 = $18. The answer is $18.

Done

Diffusion LLM: many words per step

8B parameters · 612 tokens/s

Janet starts with 16 eggs a day. She eats 3 and bakes with 4, so 16 − 3 − 4 = 9 eggs are left. She sells them for $2 each: 9 × $2 = $18. The answer is $18.

Done, 2.7× faster

ANNUAL DECODING SAVINGS
at DeepSeek's 168B tokens/day

$6.1M	$9.2M	$11.5M
if 1.5× output per GPU	if 2× output per GPU	if 2.7×, as in the demo

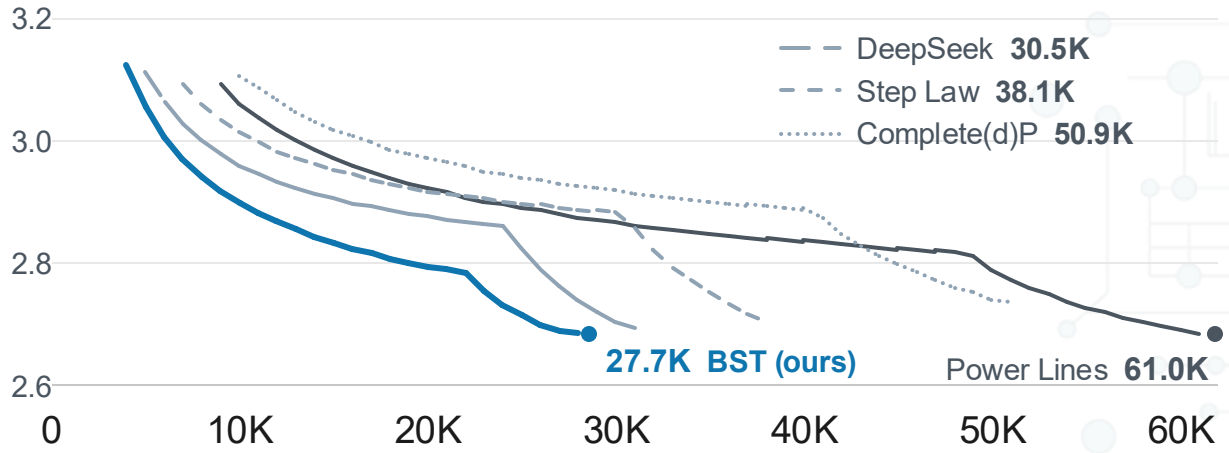
Illustrative replay, slowed for display. Savings = 168B output tokens/day × $0.30 per 1M decoding cost (DeepSeek, Feb 2025; H800 at $2/GPU-hour) × (1 − 1/s), s = diffusion's output per GPU under load; the demo's 2.7× is single-stream. Sources: Fang, Li et al., Amazon AGI (2026), arXiv 2610.04198; DeepSeek Open Infra Index (2025).

TRAINING GAINS

First-principles scaling: same loss, 55% fewer iterations

Validation loss during training, 1.4B-parameter model

Same 20B tokens for every rule · lower is better · iterations in thousands



Source: AdamW Scaling Laws from First Principles (2026, under review), Fig. 1 left, App. C.3, Table 1. [1]Step Law (Li et al., 2025): 3,700 runs, ~1M H800 GPU-hours.

−55%

iterations to match the best published rule, Power Lines

No fitting runs

others need up to ~1M GPU-hours[1] to fit

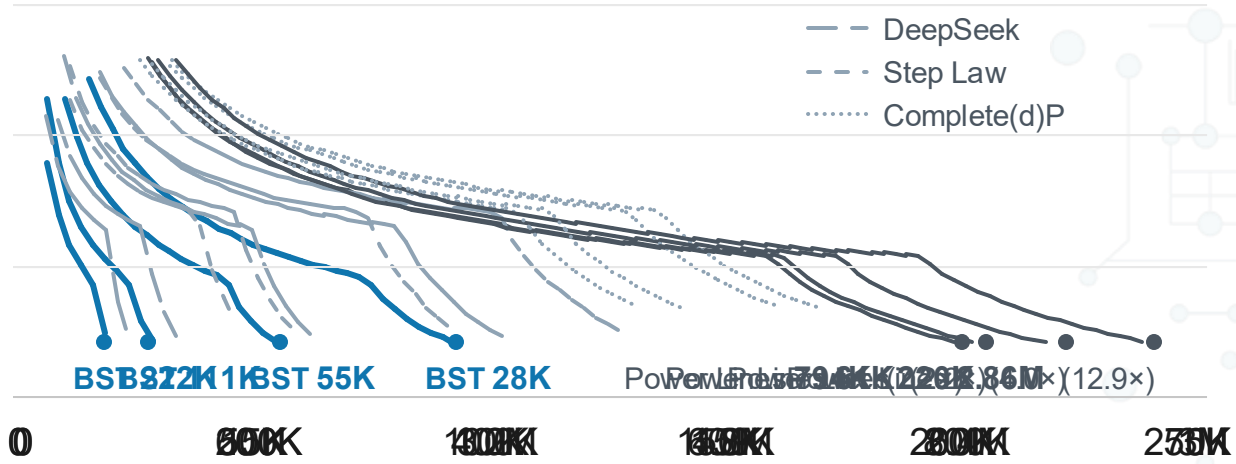
TRAINING GAINS

The iteration gap grows with model size: 13× at 512×

Validation loss vs iterations, projected to larger models

Same tokens per parameter · curve shapes from the 1.4B run

512× · 717B parameters, 10T tokens



Iterations to train, thousands

Model size multiple and parameters

Rule	1× 1.4B	8× 11B	64× 90B	512× 717B
BST (ours)	**28**	**55**	**111**	**222**
DeepSeek	31	63	129	264
Step Law	38	93	227	554
Complete(d)P	51	163	523	1,675
Power Lines	61	220	794	2,864

COST SAVED VS STEP LAW
if the cluster is held for the run

$10K	$1.0M	$93M
8× · 11B model	64× · 90B model	512× · 717B model

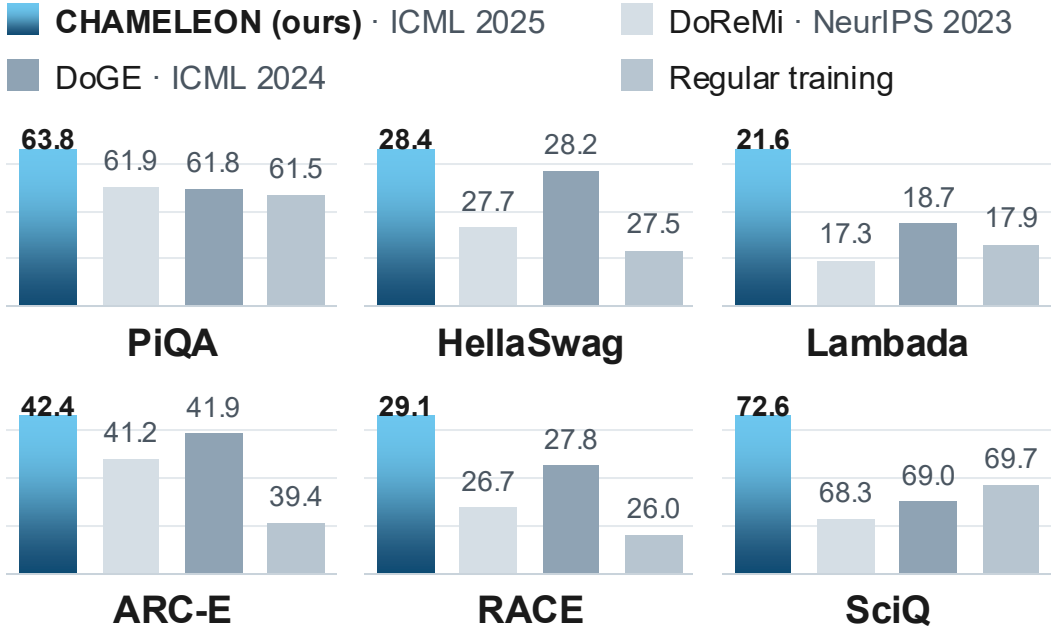
Projection: iterations = tokens ÷ batch, using each rule's batch-size law (Table 1) at 14 tokens per parameter. Cost: 6·N·T FLOPs, H100-class GPUs at 40% utilisation, $2/GPU-hour; saving = Step Law's extra iterations at BST's cost per iteration. Without a cluster limit, GPU-hours are similar and BST finishes sooner.

PERFORMANCE GAINS

Better data and looping both raise benchmark scores

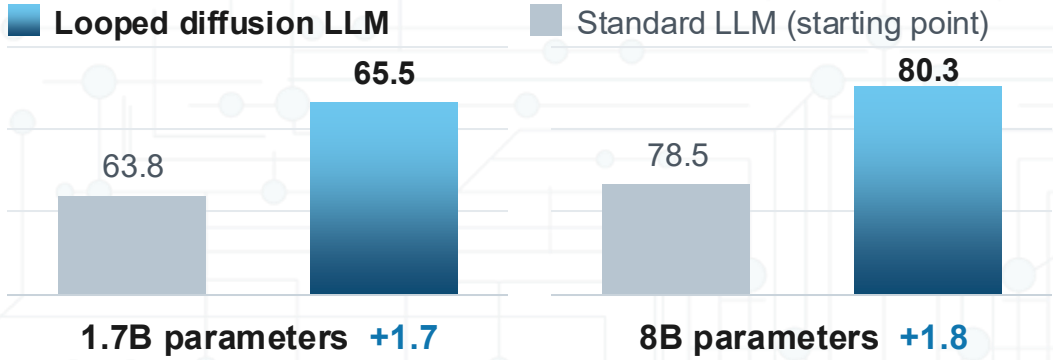
Data mixing: reasoning scores, 1.2B model

Higher is better · each axis starts above zero



Looping: average score on 11 benchmarks

Looped diffusion LLM vs the standard LLM it was built from



VS. A DIFFUSION LLM WITHOUT LOOPING, SAME ACCURACY

−13.6%
compute per token

+8.5%
answers per second

Sources: W. Xie, F. Tonin, V. Cevher, Chameleon (ICML 2025); Fang, Li et al., Amazon AGI (2026), arXiv 2610.04198. Compute comparison on GSM8K at 93.3% accuracy.

WHAT YOU GAIN

Cheaper to serve, cheaper to train, and a better model

SERVING

$11.5M

decoding cost saved per year at DeepSeek scale, at most

2.7× faster than a standard LLM on the same GPU

TRAINING

$2M

fitting compute saved per rule

2.2× fewer iterations than the best published rule

PERFORMANCE

+3.7 points

best per-benchmark gain from data mixing vs regular training

Looping: 13.6% less compute at the same accuracy

EPFL AI Center

ETH AI CENTER

OUR APPROACH

Our edge: science-first scaling

Theory first: derive the scaling rules on small models, then train big.

01

Convert

Turn Swiss open LLM Apertus into a diffusion LLM.

02

Scale with theory

BST rules carry us to ~8B.

03

Post-train & align

Helpful, safe, reliable.

04

Adapt compute

Looping spends compute where prompts are hard.

12-month roadmap

Months 1–6	Months 7–9	Months 8–12
Convert Apertus, set the recipe on small models	**Train ~8B**	**Post-train & release**

EPFL AI Center

ETH AI CENTER

OUTCOMES, IMPACT & COLLABORATION

What you get, where it pays off, and how to join

01 WHAT YOU GET

A fast Swiss LLM
~8B, release planned

A scaling playbook
rules derived from theory

Adjustable compute
more compute on hard tasks

02 WHERE IT PAYS OFF

Customer service & assistants
faster, cheaper replies

Software & code tools
fast code completion

Banking, pharma & public sector
open and self-hosted

03 HOW TO JOIN

Pilot partner
test it on your use case

Co-development
a model tuned to your data

Compute partner
hardware and deployment

EPFL AI Center

ETH AI CENTER

TRACK RECORD

Built on our peer-reviewed research in top-tier ML venues

SCALING & OPTIMIZATION THEORY

ICML 2026 — **On the Role of Batch Size in Stochastic Conditional Gradient Methods**
R. Islamov, …, A. Lucchi, …, V. Cevher

ICML 2025 — **Training Deep Learning Models with Norm-Constrained LMOs**
T. Pethick, …, V. Cevher

ICLR 2025 — **Adaptive Methods through the Lens of SDEs: Theoretical Insights on the Role of Noise**
E. M. Compagnoni, …, A. Lucchi

NeurIPS 2024 — **On Feature Learning in Structured State Space Models**
L. C. Vankadara, J. Xu, M. Haas, V. Cevher

Under review — **AdamW Scaling Laws from First Principles**
2026 · source of the BST results in this deck

DATA & FAST GENERATION

ICML 2025 — **Chameleon: A Flexible Data-mixing Framework for Language Model Pretraining and Finetuning**
W. Xie, F. Tonin, V. Cevher

ICLR 2025 — **Faster Inference of Flow-Based Generative Models via Improved Data-Noise Coupling**
A. Davtyan, L. Tadesse Dadi, V. Cevher, P. Favaro

TEST-TIME COMPUTE & ALIGNMENT

ICLR 2025 — **Efficiently Learning at Test-Time: Active Fine-Tuning of LLMs**
J. Hübotter, S. Bongni, I. Hakimi, A. Krause

NeurIPS 2024 — **Group Robust Preference Optimization in Reward-free RLHF**
S. S. Ramesh, …, I. Bogunovic

EPFL AI Center
ETH AI CENTER

Thanks !

EPFL AI Center

ETH AI CENTER

Become a pilot partner

Build faster, Swiss-made AI with us.

Talk to us after the session →

EPFL · ETH Zurich · University of Basel · University of Bern