Is More Data Worth the Cost? Dataset Scaling Laws in a Tiny Attention-Only Decoder

Götz-Henrik Wiegand[1] , Lorena Raichle[1], Rico Städeli[1], Tomas Hrycej[1], Bernhard Bermeitinger[2] and Siegfried Handschuh[1]

[1]Institute of Computer Science, University of St. Gallen (HSG), St. Gallen (CH) | [2]Institute of Computer Science Vorarlberg, University of St. Gallen (HSG), Dornbirn (AT)

DATA-FM Workshop @ ICLR 2026, Rio de Janeiro, Brazil

University of St.Gallen
Institute of Computer Science

Paper

Motivation

Problem & Related Work: Training LLMs requires balancing…

Dataset | **Compute** | **Model**

- More data improves performance, but at increasing cost!
- Scaling Laws predict diminishing returns (Kaplan et al., 2020)

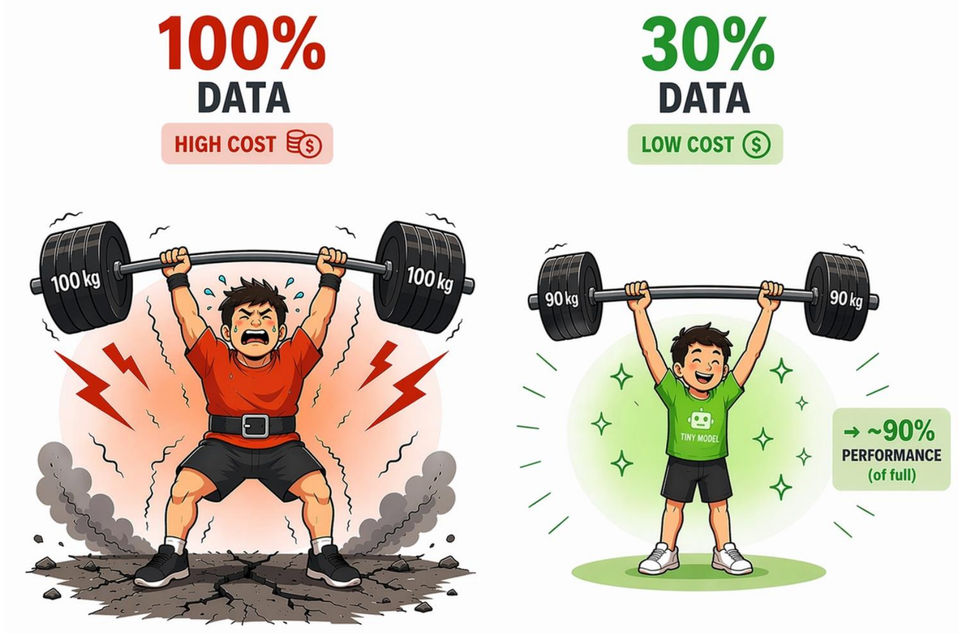
When is more data no longer worth the cost?

Focus: Tiny, compute-constrained models

Research Questions:

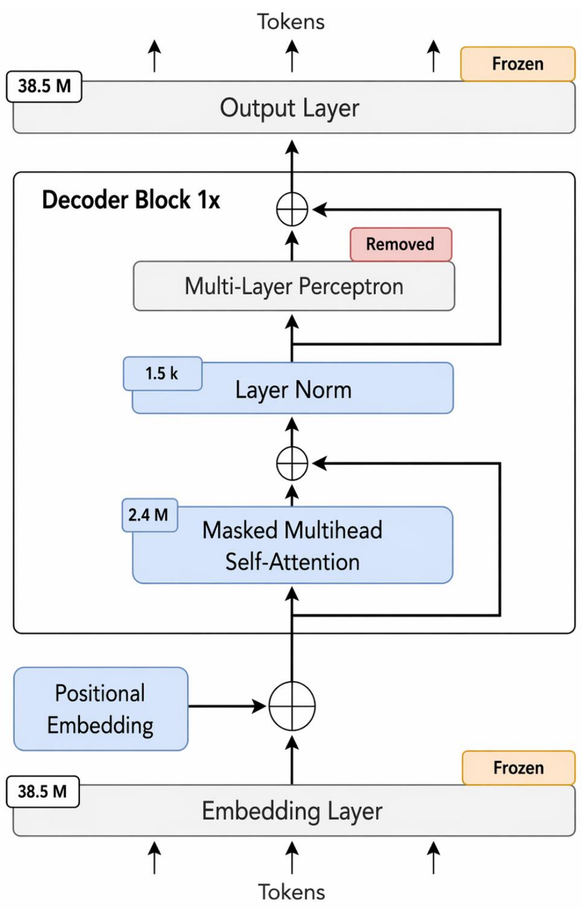
- Do diminishing returns persist in tiny models?
- How much data is needed for near-full performance?
- What is the cost–performance trade-off of dataset scaling?

Key Insight:



Tiny Models achieve 90% of performance with only ~30% of data!

Model Architecture



Controlled Setup:

Frozen embeddings

Frozen output layer

Attention-Only Decoder:

Removed MLP

Parameters:

~ 2.4M trainable params

~ 80M total params

Architecture:

- 1 Decoder
- 1 Head
- 768 Embedding

Freezing representations to isolate data scaling in self-attention.

Subset Construction

- *AllTheNews2.0* dataset (validation experiments on *Wikitext-103*)
- Nested power-of-two subsets ($2^7 \rightarrow 2^{17}$), each extending the previous subset
- Fixed seq. ordering, fixed seq. length of 1024 tokens
- Moderate subsets approximate full dataset distribution

Subset 2^7: [■■□□□□□□□□□□□□□□□]
…
Subset 2^9: [■■■■■■□□□□□□□□□□□]
…
Subset 2^{13}: [■■■■■■■■■■■■□□□□]
…
Full Dataset 2^{17}: [■■■■■■■■■■■■■■■■■]

Cost-Performance Trade-Off

Under fixed Training Schedule:

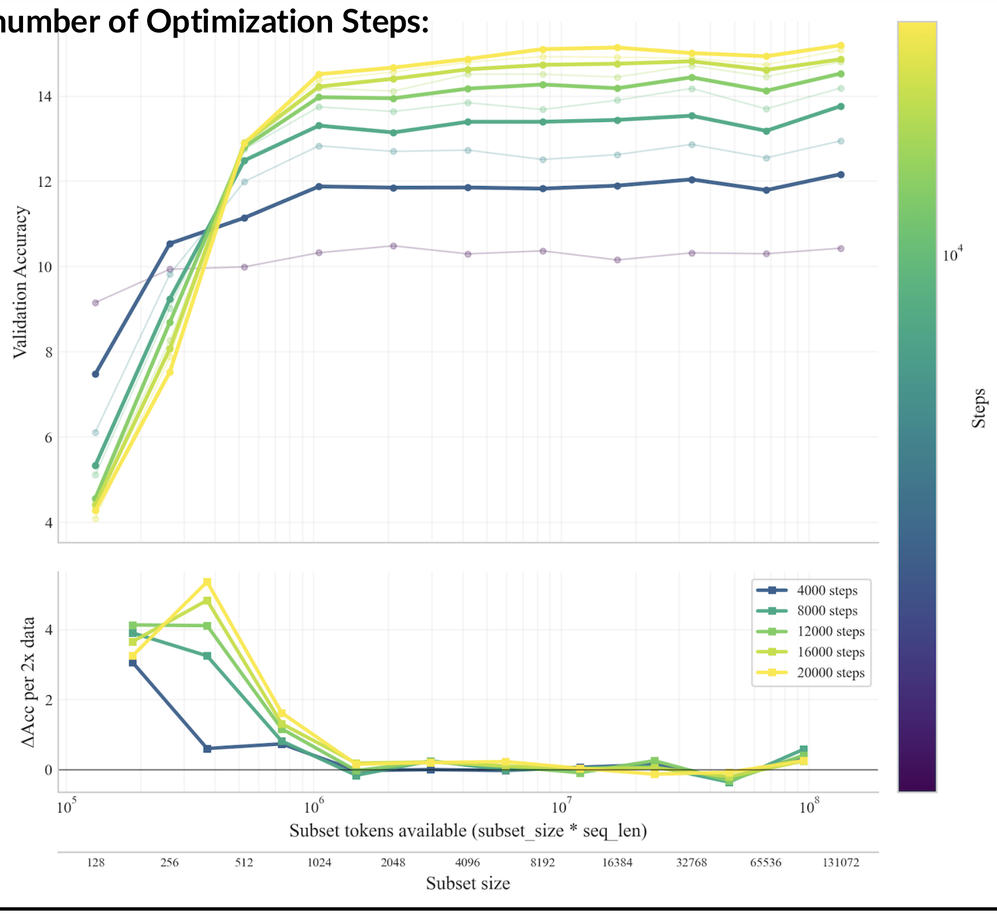
Fraction of full-data Accuracy	Subset	Data (%)	Train Time (h)	Cost (%)
80%	2^{13}	7	8	7
90%	**2^{15}**	**29**	**32**	**26**
95%	2^{16}	56	60	49
100%	2^{17}	100	123	100

Dataset scaling shows **early saturation**:

- most performance is achieved with moderate data
- additional data = limited gains at increasing cost

Results and Conclusion

Fixed number of Optimization Steps:



Key Takeaways:

- Performance gains saturate while cost grows nearly linearly
- Additional data alone is insufficient without more compute
- Moderate data regimes offer best trade-off

→ Scaling Laws remain actionable in tiny architectures & low-resource settings

→ Data Efficiency is achievable in tiny models

Practical Impact:

- Small fraction of data is often sufficient for near-full performance
- Trade-Offs reduce compute cost for small labs and rapid experimentation
- Enables efficient pretraining under limited budgets

Data–compute trade-offs enable near-optimal performance at significantly reduced cost!