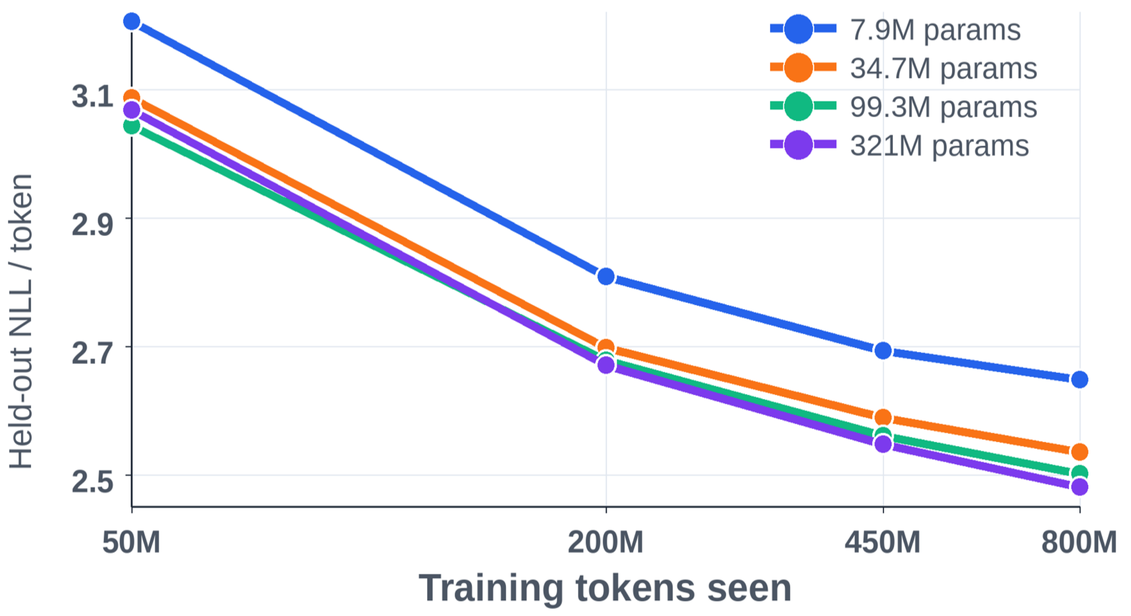
More tokens reduce loss, but bigger models do not win at fixed compute.

Scaling laws for autoregressive foundation models on observational health event streams

Background: We tested whether autoregressive foundation models trained on 2.9M IPCI patient timelines show predictable scaling under deterministic held-out evaluation.

Within the tested range, more training data continued to reduce held-out loss.

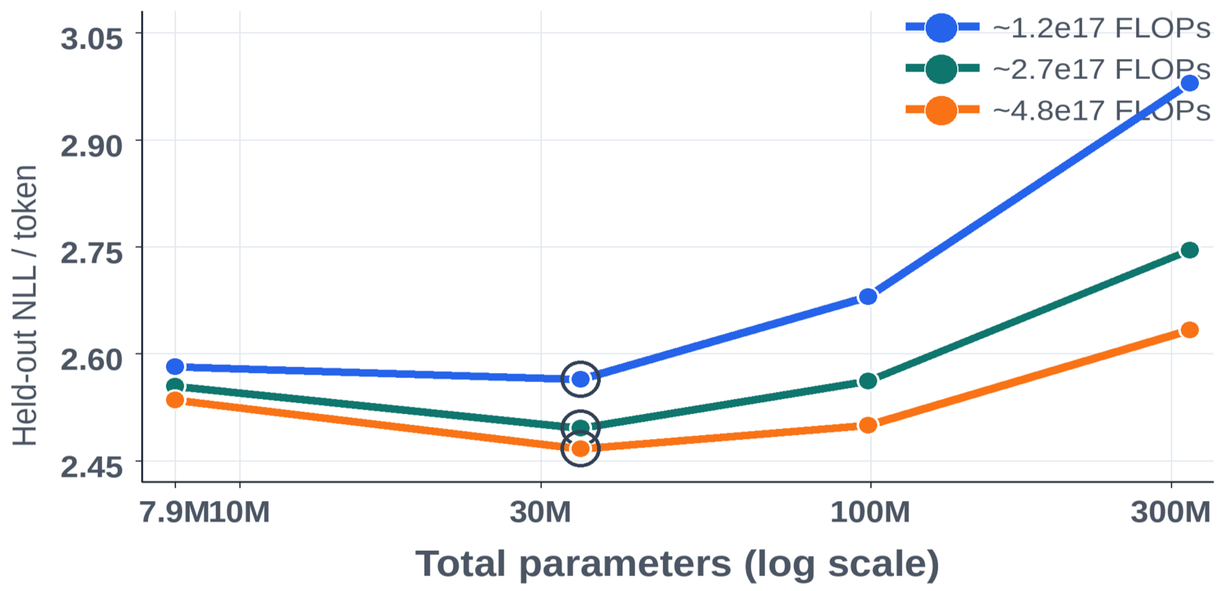
2.9M IPCI patients • 4 model sizes • 50M to 800M training tokens



More training data kept improving loss, indicating pretraining had not yet saturated.

At fixed compute, balancing model size and token budget beat scaling parameters alone.

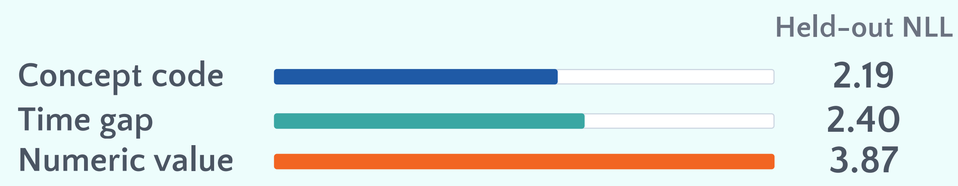
Matched compute budgets on the same IPCI training data



Balanced model size and token budget beat scaling parameters alone at fixed compute.

Value tokens remained hardest to predict

Value tokens encode numeric measurements such as labs and vitals.



Scaling setup

DATA — Prepare event streams

2.9M IPCI patients
Patient-disjoint token sequences
Longitudinal OMOP records

AR — Train model families

Decoder-only AR models
Data-scaling runs
Matched-compute runs

NLL — Evaluate held-out loss

Deterministic next-token NLL
Compare model sizes
Concept, time, value tokens

Limitation: These results are based on deterministic held-out next-token loss in IPCI event streams. Future work should test whether the same scaling patterns hold across other datasets, architectures, and downstream tasks.

OHDSI

Egill A. Fridgeirsson[1], Peter Rijnbeek[1]
e.fridgeirsson@erasmusmc.nl
[1] Medical Informatics, Erasmus University Medical Center

Erasmus MC University Medical Center Rotterdam

Health Data Science