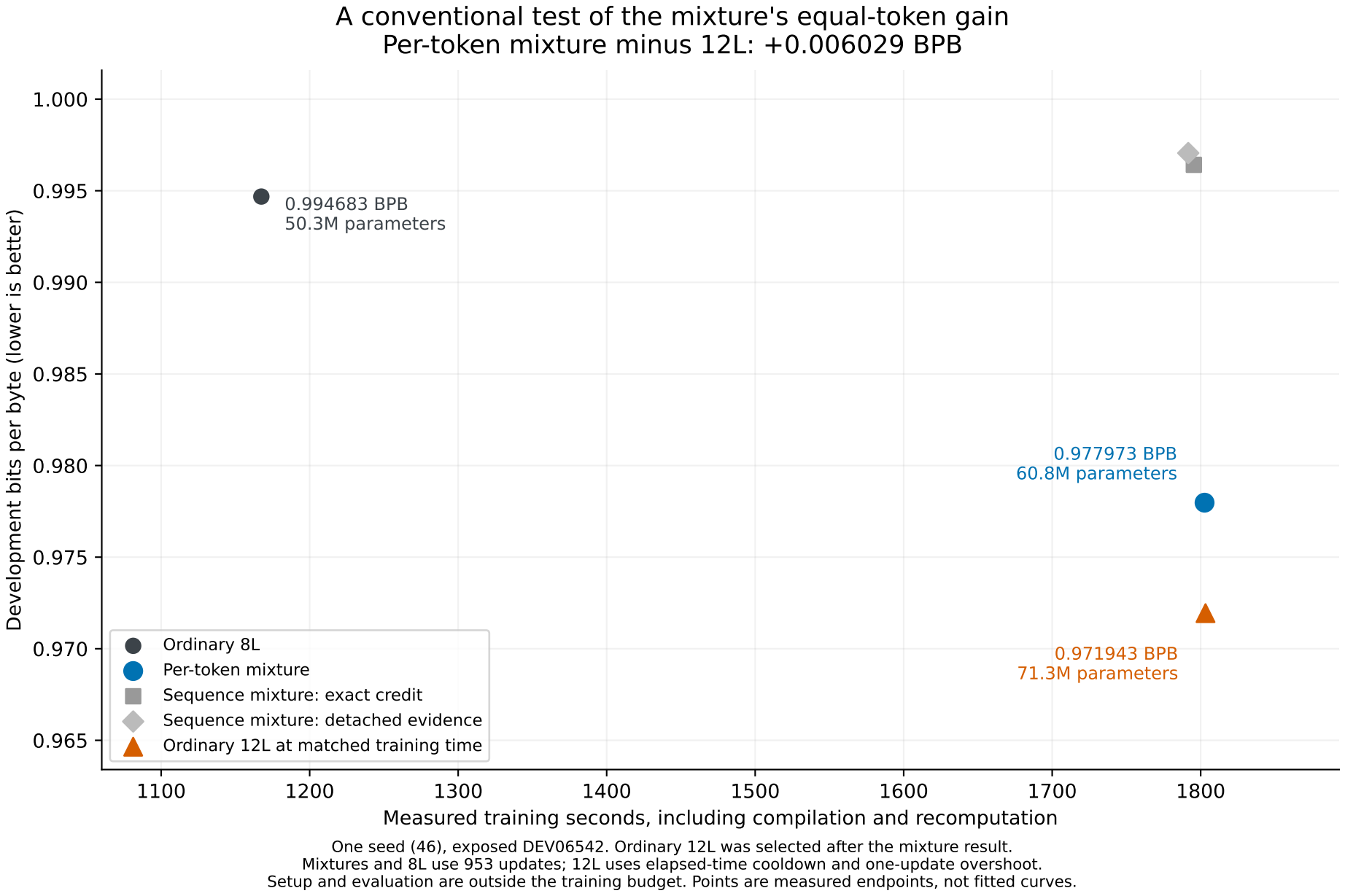
A conventional test of the mixture's equal-token gain
Per-token mixture minus 12L: +0.006029 BPB
Development bits per byte (lower is better)
1.000
0.995
0.990
0.985
0.980
0.975
0.970
0.965
0.994683 BPB
50.3M parameters
0.977973 BPB
60.8M parameters
0.971943 BPB
71.3M parameters
Ordinary 8L
Per-token mixture
Sequence mixture: exact credit
Sequence mixture: detached evidence
Ordinary 12L at matched training time
1100
1200
1300
1400
1500
1600
1700
1800
Measured training seconds, including compilation and recomputation
One seed (46), exposed DEV06542. Ordinary 12L was selected after the mixture result.
Mixtures and 8L use 953 updates; 12L uses elapsed-time cooldown and one-update overshoot.
Setup and evaluation are outside the training budget. Points are measured endpoints, not fitted curves.