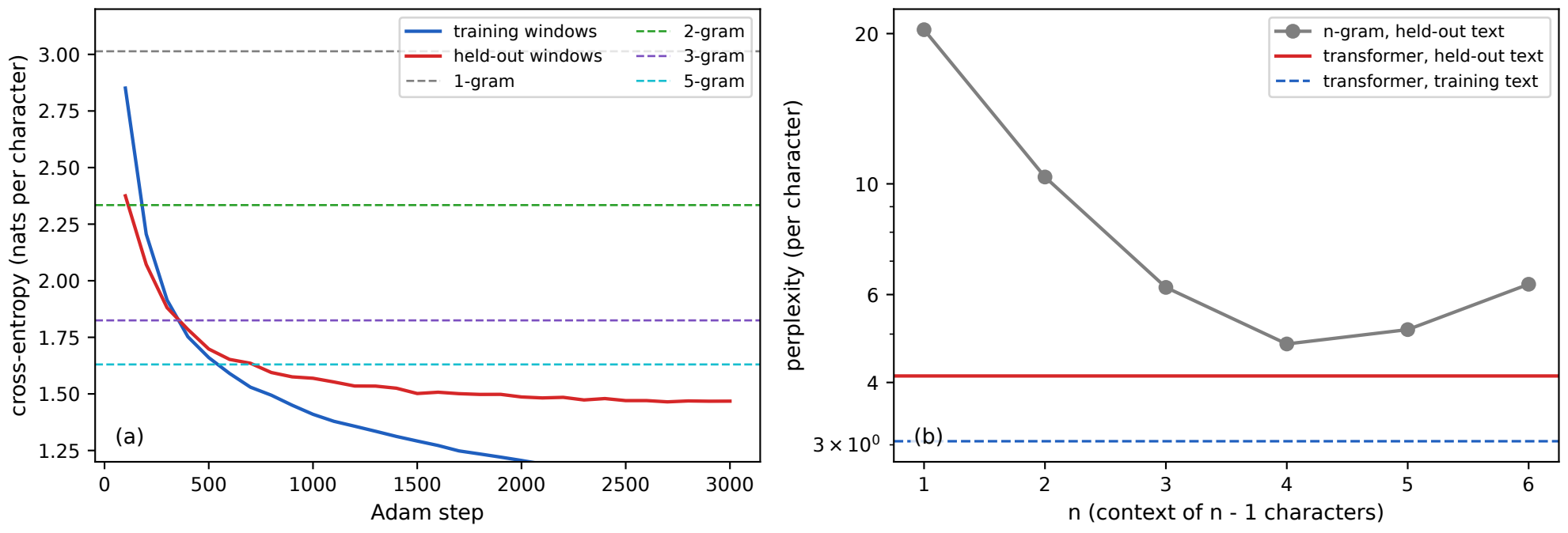

training windows
held-out windows
1-gram
2-gram
3-gram
5-gram
cross-entropy (nats per character)
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
0
500
1000
1500
2000
2500
3000
Adam step
(a)
n-gram, held-out text
transformer, held-out text
transformer, training text
perplexity (per character)
20
10
6
4
3 × 10^0
1
2
3
4
5
6
n (context of n - 1 characters)
(b)