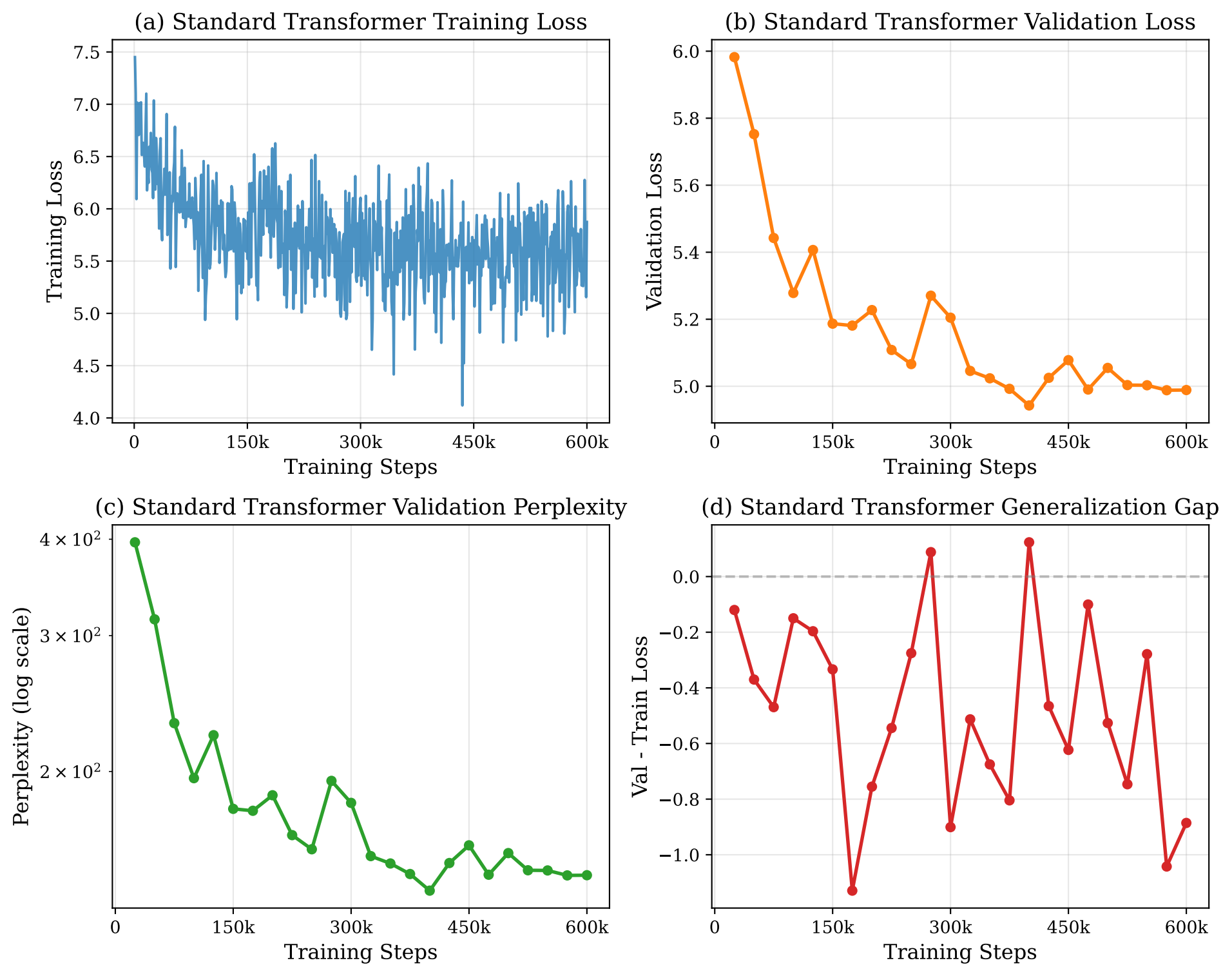
(a) Standard Transformer Training Loss
Training Loss
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
0
150k
300k
450k
600k
Training Steps
(b) Standard Transformer Validation Loss
Validation Loss
6.0
5.8
5.6
5.4
5.2
5.0
0
150k
300k
450k
600k
Training Steps
(c) Standard Transformer Validation Perplexity
Perplexity (log scale)
4 × 10²
3 × 10²
2 × 10²
0
150k
300k
450k
600k
Training Steps
(d) Standard Transformer Generalization Gap
Val - Train Loss
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
0
150k
300k
450k
600k
Training Steps