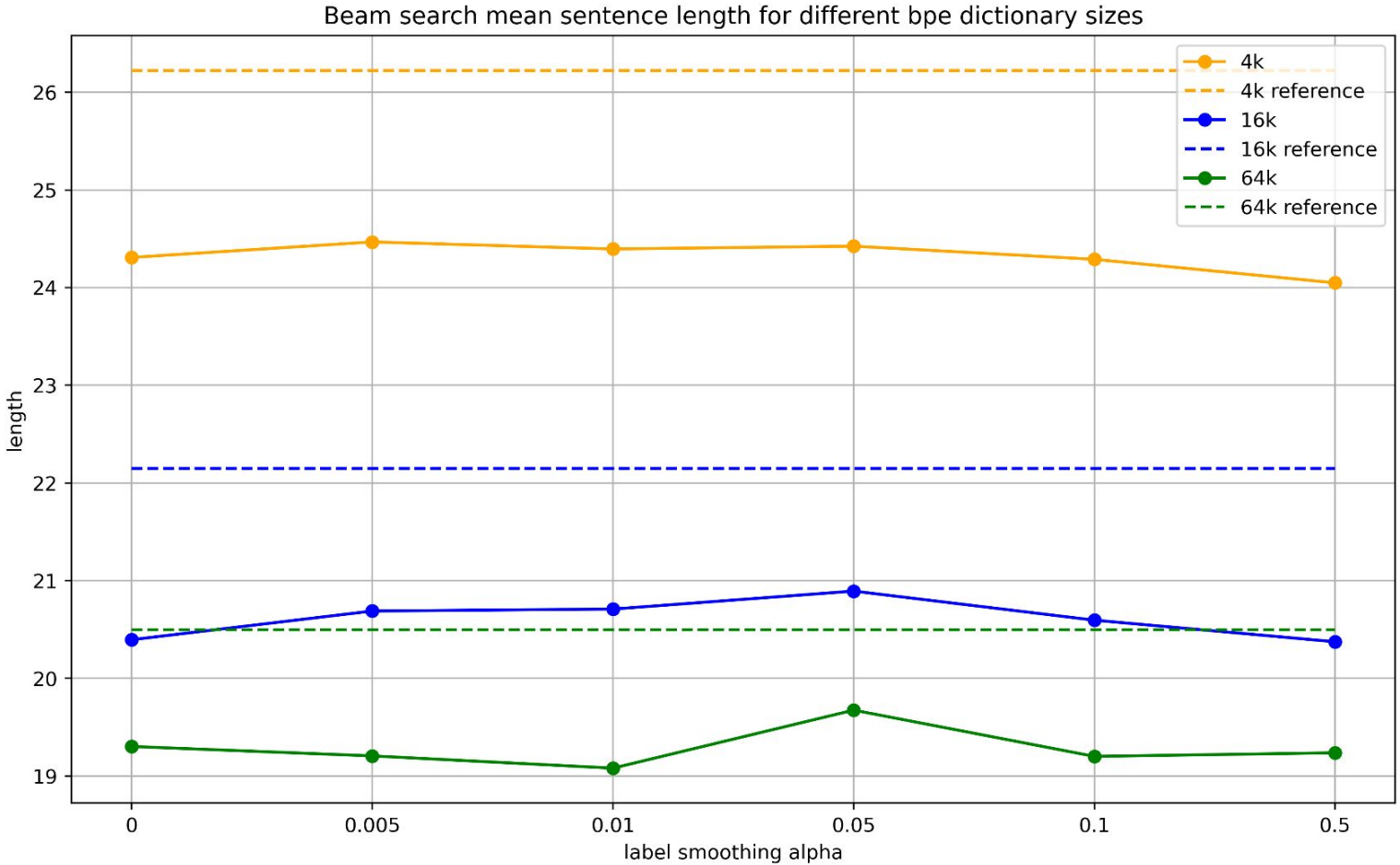
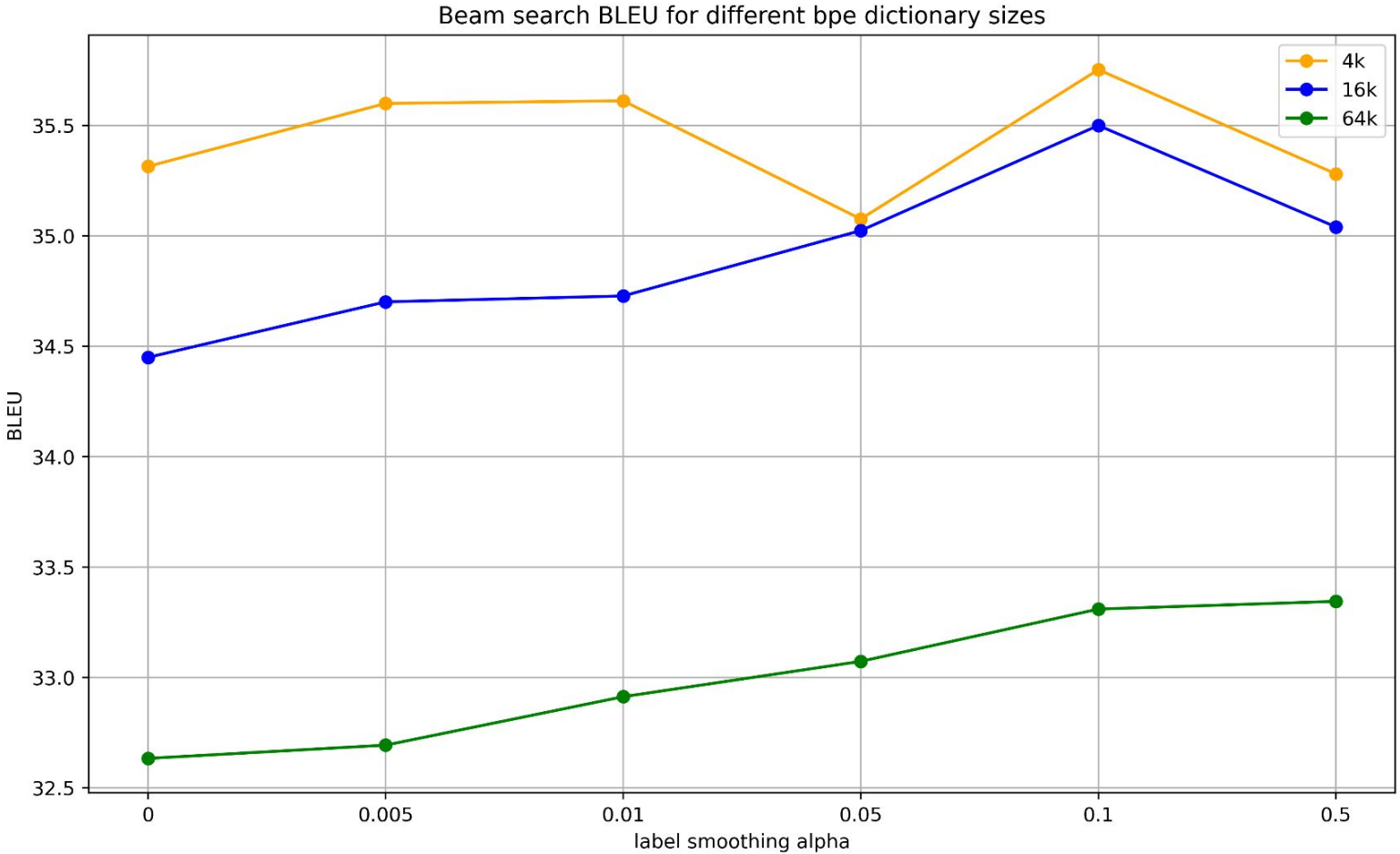
Transformer experiments

Dict Size Experiments - Beam Search

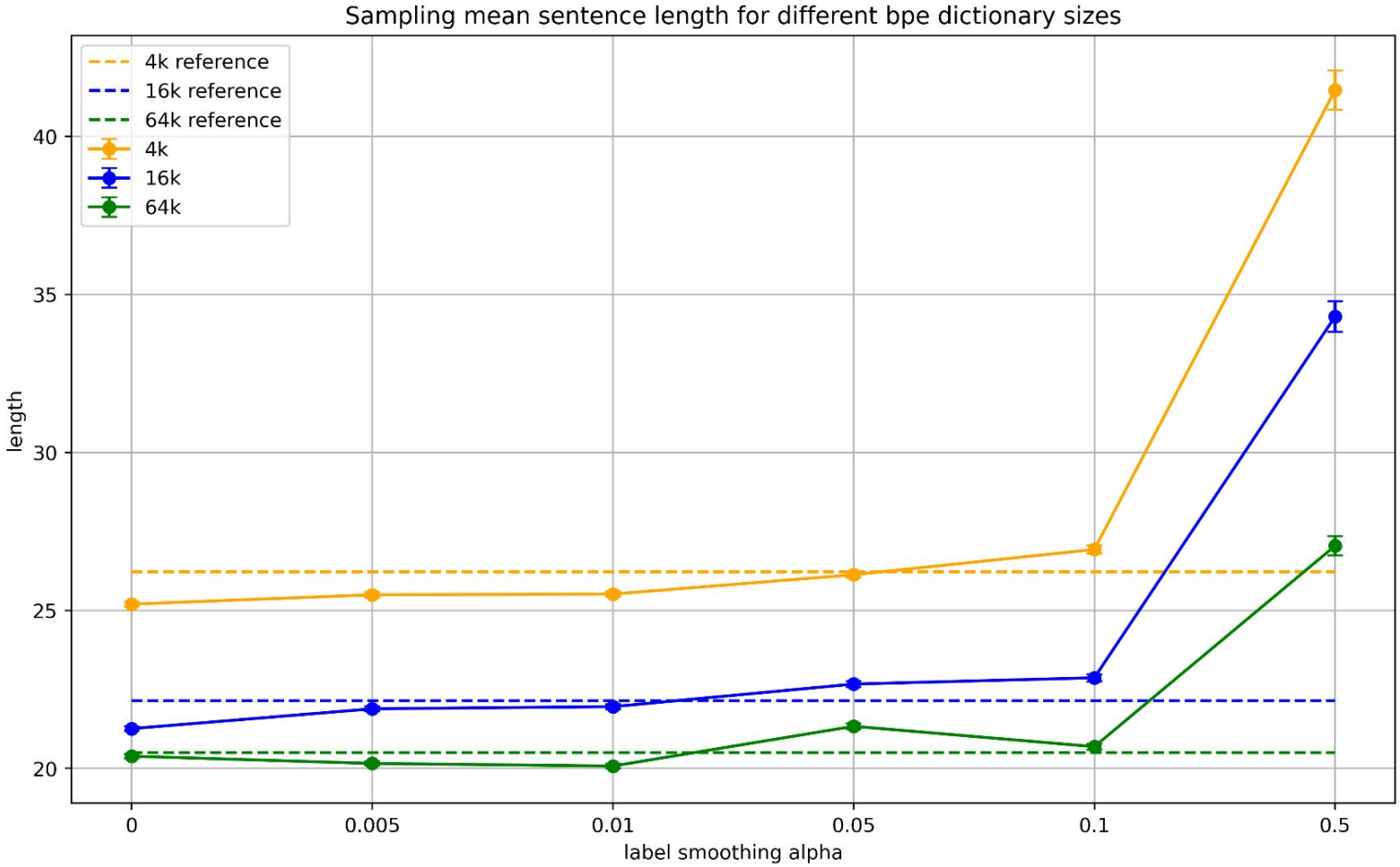
- Transformer setup as in Riley and Chiang
  - 6 layers, 4 heads
  - fairseq transformer_iwslt_de_en setup
  - sinusoid positional embeddings
  - context size of 4096 tokens
  - Trained on iwslt17 de-en data
- Experimental setup
  - Evaluation with beam size of 4
  - We compare the mean sentence length of the beam search outputs to the mean sentence length of the reference translations
    - sentence length = # bpe tokens
    - evaluated for label smoothing alpha values [0, 0.005, 0.01, 0.05, 0.1, 0.5] and bpe dict sizes [4k, 16k, 64k]
  - BLEU values reported for reference


Beam search mean sentence length for different bpe dictionary sizes
4k
4k reference
16k
16k reference
64k
64k reference
length
26
25
24
23
22
21
20
19
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha


Beam search BLEU for different bpe dictionary sizes
4k
16k
64k
BLEU
35.5
35.0
34.5
34.0
33.5
33.0
32.5
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha

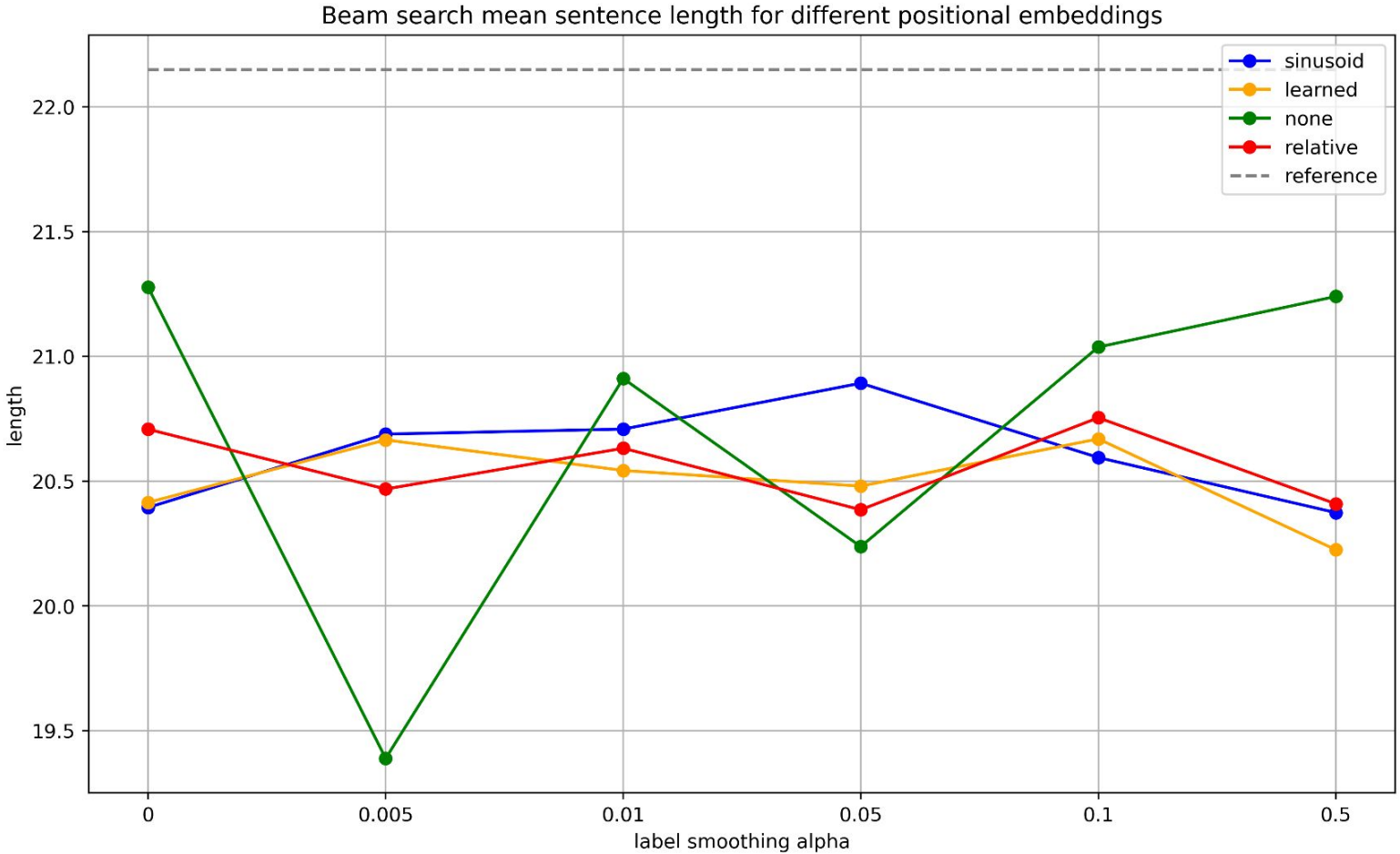
Dict Size Experiments - Sampling

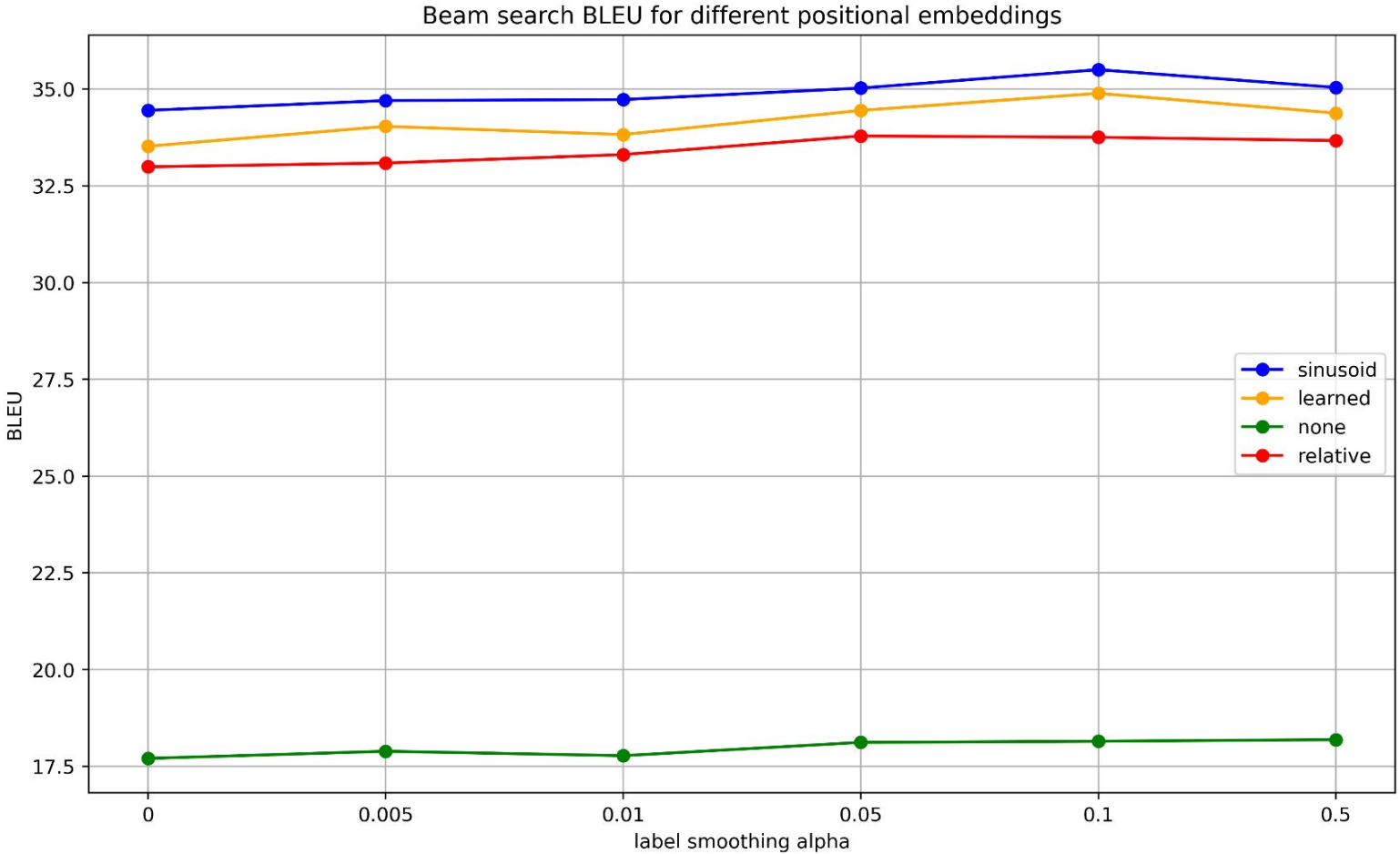
- Sample 1000 hypotheses per source sentence
- We compute the mean hypothesis length per sentence and then compute the mean of means for each model
- We compare it again to the mean length of the reference corpus

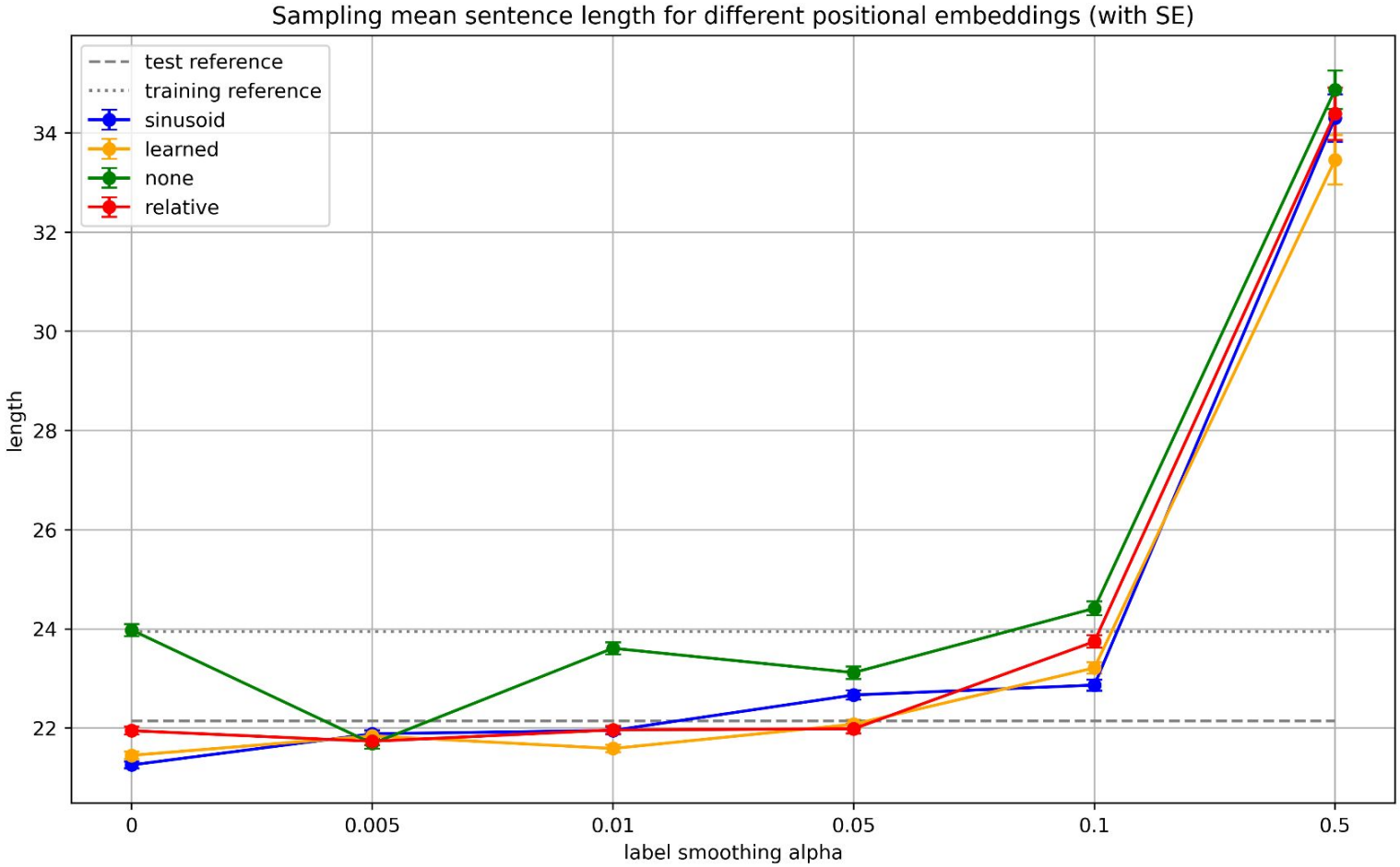

Sampling mean sentence length for different bpe dictionary sizes
4k reference
16k reference
64k reference
4k
16k
64k
length
label smoothing alpha
0
0.005
0.01
0.05
0.1
0.5
20
25
30
35
40

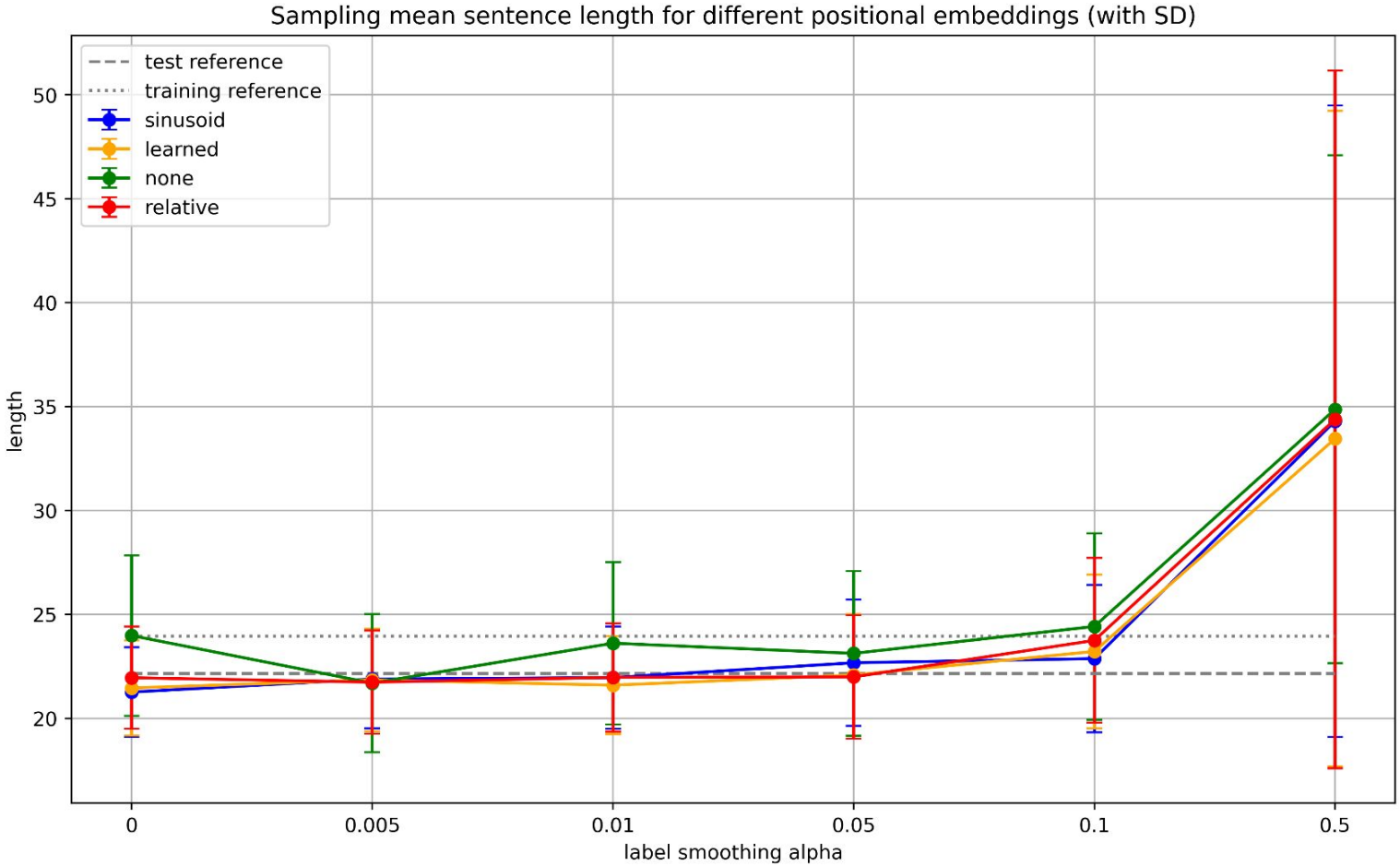
Positional Embedding Experiments

- Comparison between sinusoidal, learned, relative and no positional embeddings
  - 16k bpe dict size for all systems
- Additional comparison of sampling results to mean training set reference length (before only test set reference length)
  - We plot standard error and standard deviation for the sampling results
- Relative positional embeddings based on Attention with Linear Biases
  - Based on Florian Schottmann implementation
  - Relative positional attention in encoder and decoder self attention, no positional embeddings in cross-attention
  - Setup variation due to compatibility: 8 attention heads, fairseq version 0.10.2


Beam search mean sentence length for different positional embeddings
sinusoid
learned
none
relative
reference
22.0
21.5
21.0
20.5
20.0
19.5
length
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha

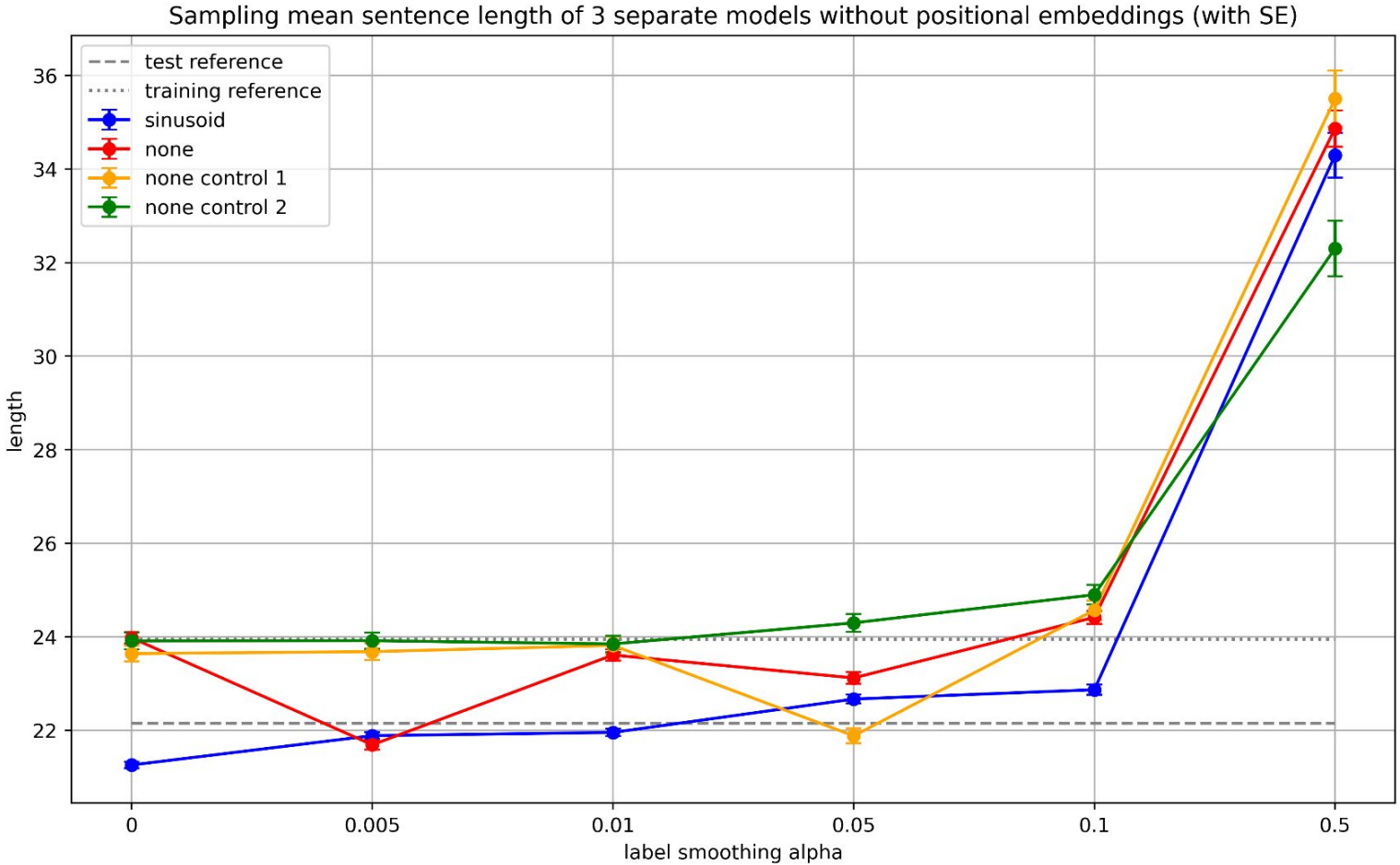

Beam search BLEU for different positional embeddings
BLEU
35.0
32.5
30.0
27.5
25.0
22.5
20.0
17.5
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha
sinusoid
learned
none
relative

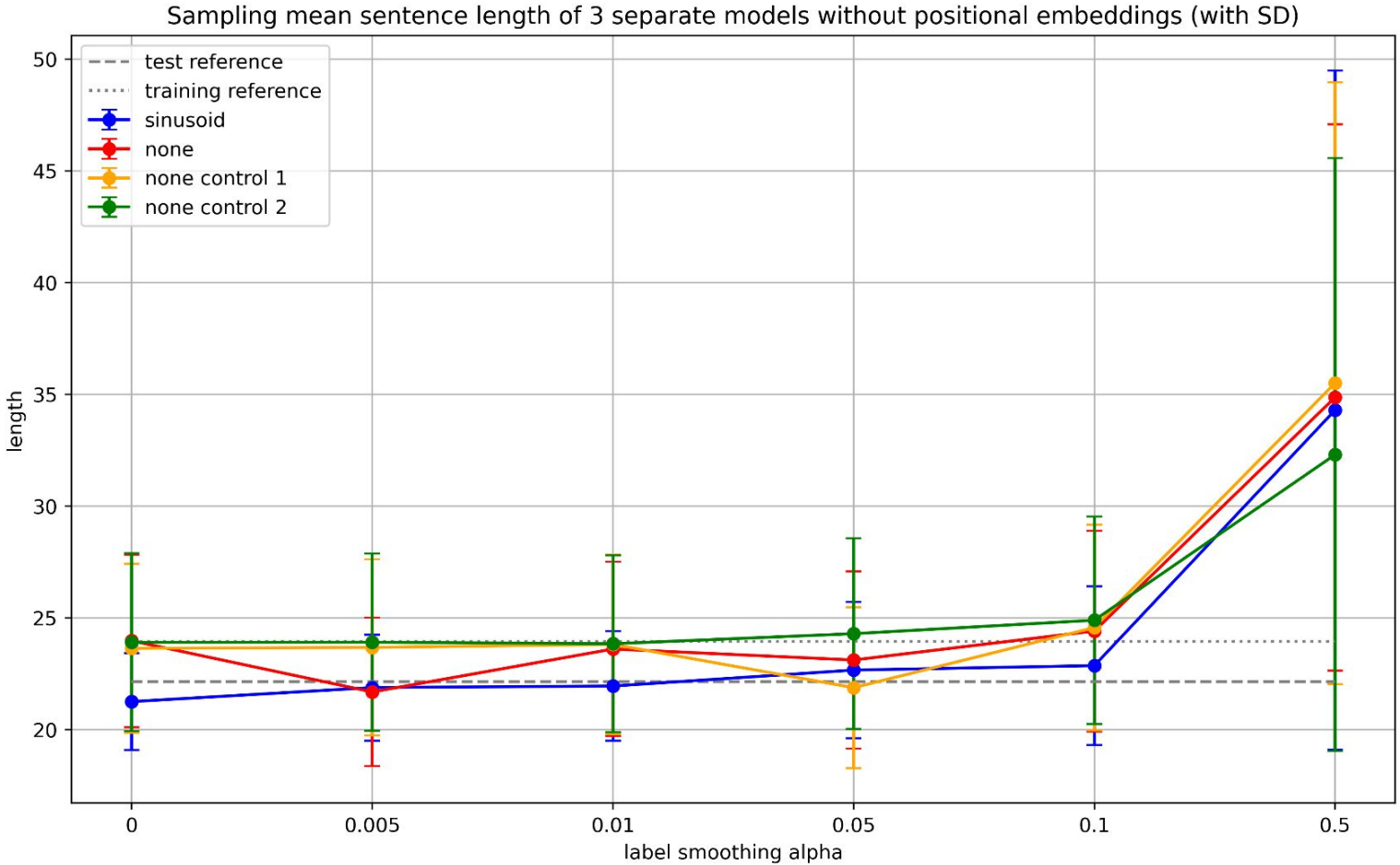

Sampling mean sentence length for different positional embeddings (with SE)
test reference
training reference
sinusoid
learned
none
relative
length
34
32
30
28
26
24
22
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha


Sampling mean sentence length for different positional embeddings (with SD)
test reference
training reference
sinusoid
learned
none
relative
length
label smoothing alpha
0
0.005
0.01
0.05
0.1
0.5
20
25
30
35
40
45
50

No Positional Embeddings Analysis

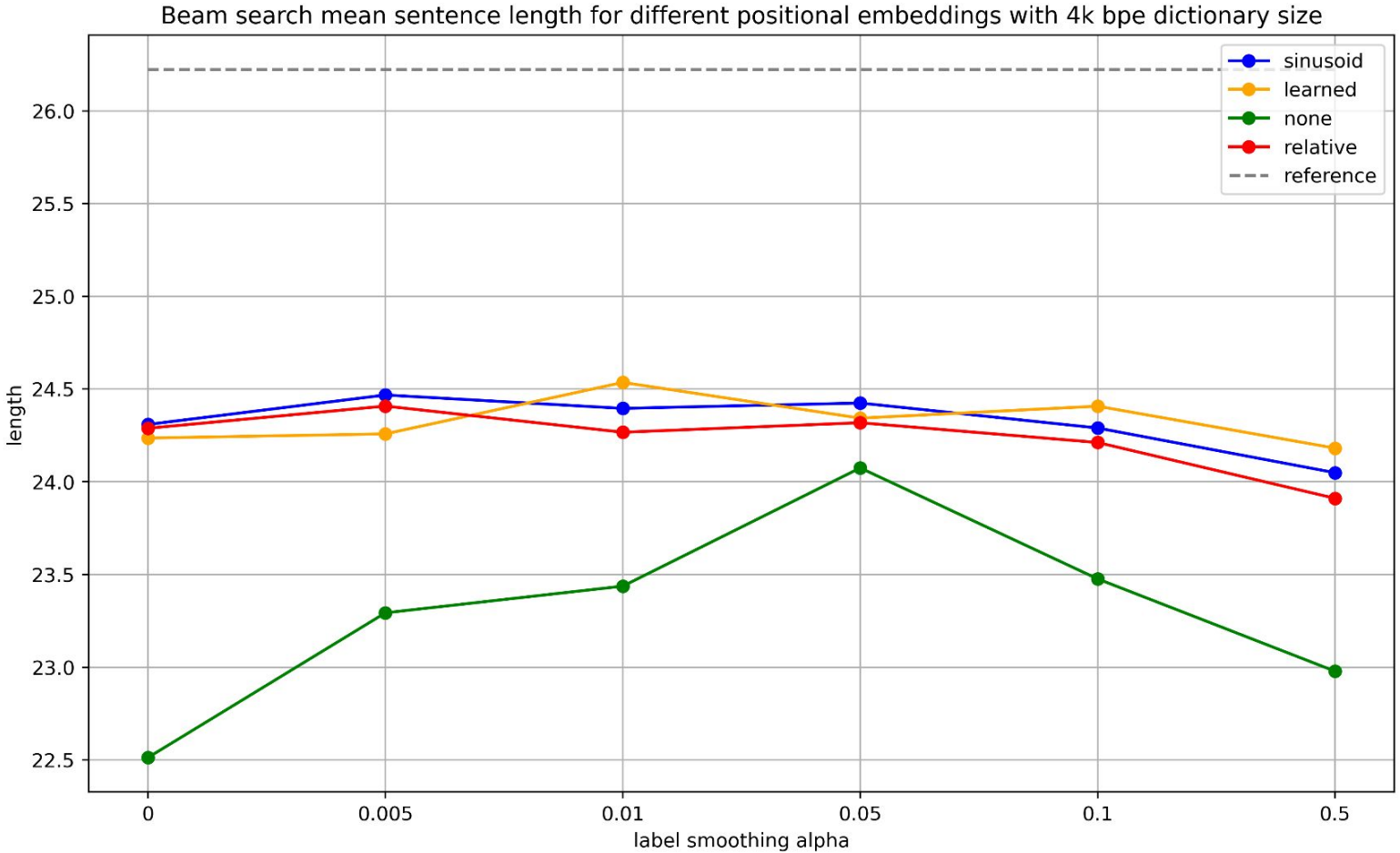
- Surprisingly good length ratio of models trained without positional embeddings
- 3 models trained without positional embeddings with different seeds for each label smoothing parameter
  - `fairseq-train` with `--no-token-positional-embeddings` flag
- Only 500 samples per sentence due to computational resources
- Observations:
  - Large variation across runs
  - Models seem to learn training set reference length well

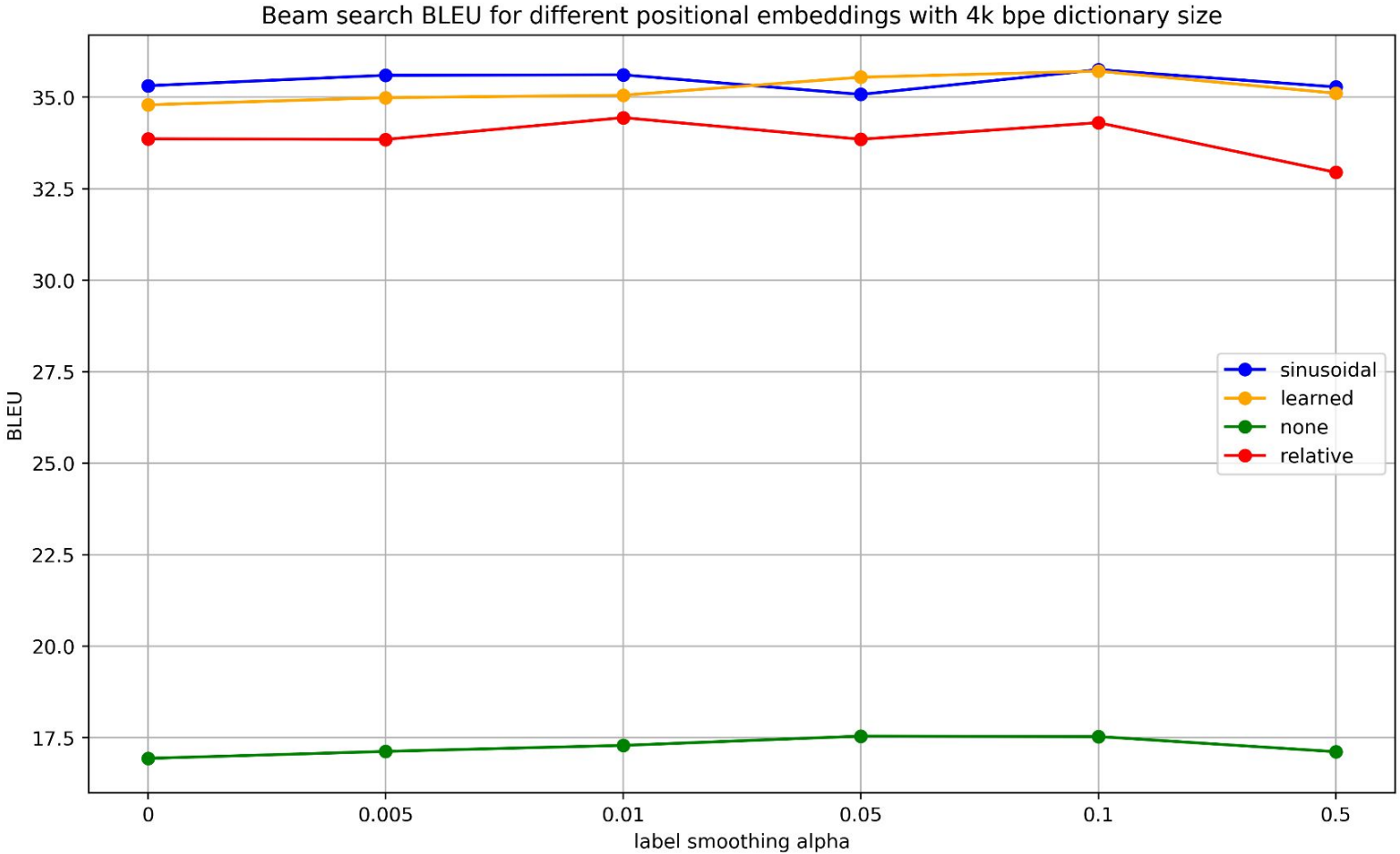

Sampling mean sentence length of 3 separate models without positional embeddings (with SE)
test reference
training reference
sinusoid
none
none control 1
none control 2
length
36
34
32
30
28
26
24
22
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha

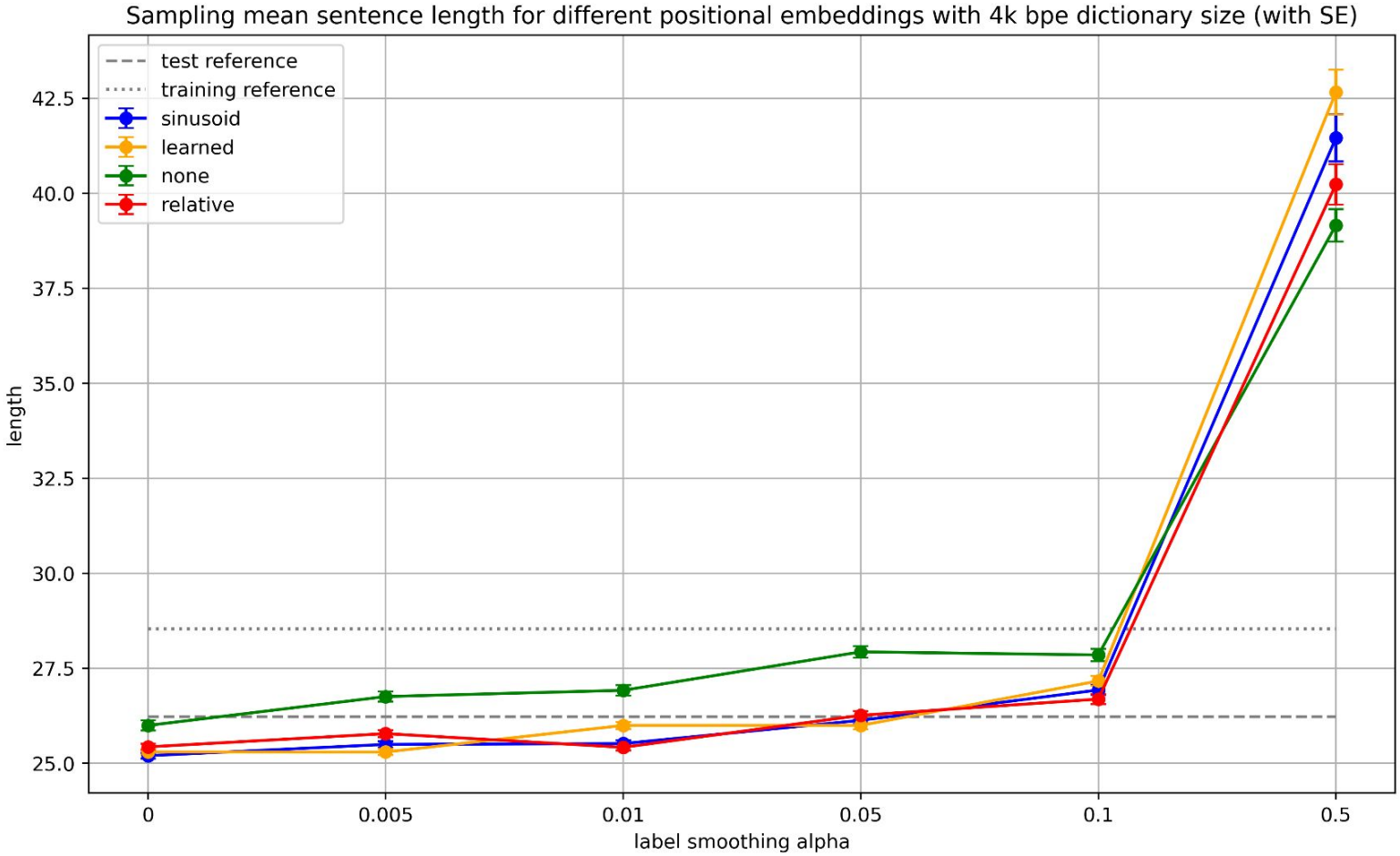

Sampling mean sentence length of 3 separate models without positional embeddings (with SD)
test reference
training reference
sinusoid
none
none control 1
none control 2
length
label smoothing alpha
50
45
40
35
30
25
20
0
0.005
0.01
0.05
0.1
0.5

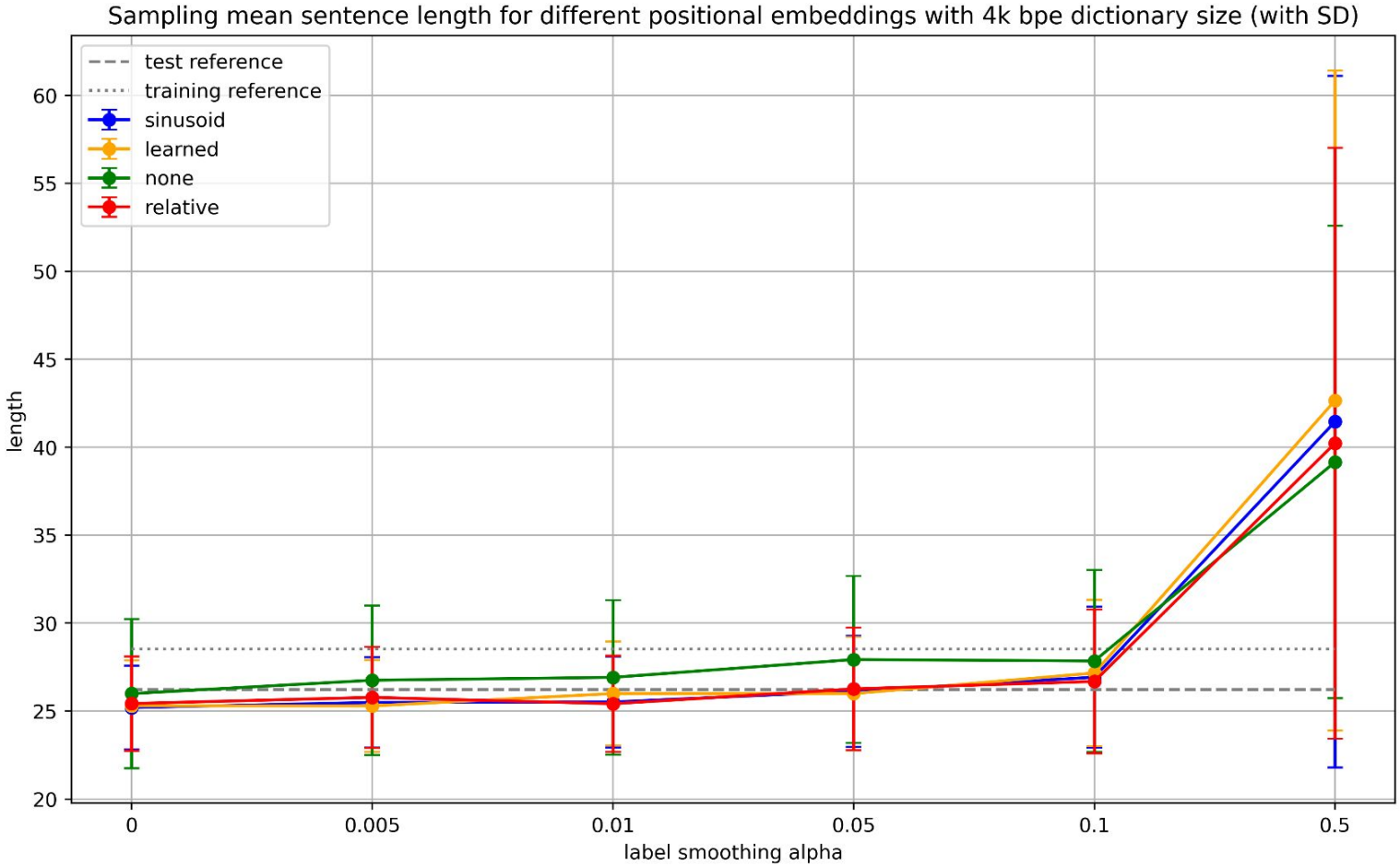
Positional embeddings dict size ablation

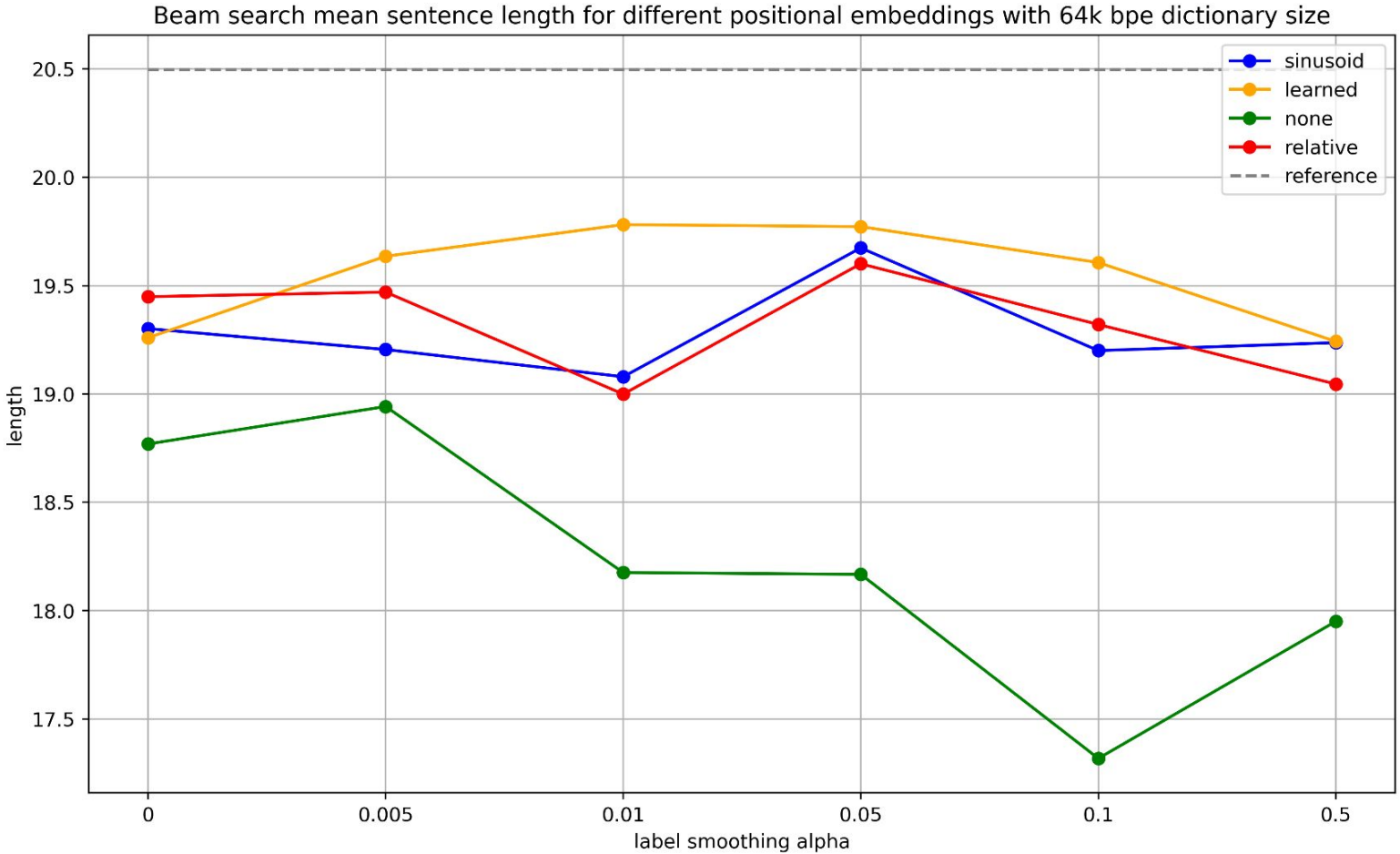
- Comparison between sinusoidal, learned, relative and no positional embeddings for bpe dict sizes 4k and 64k
- Same setup as Positional embedding experiments

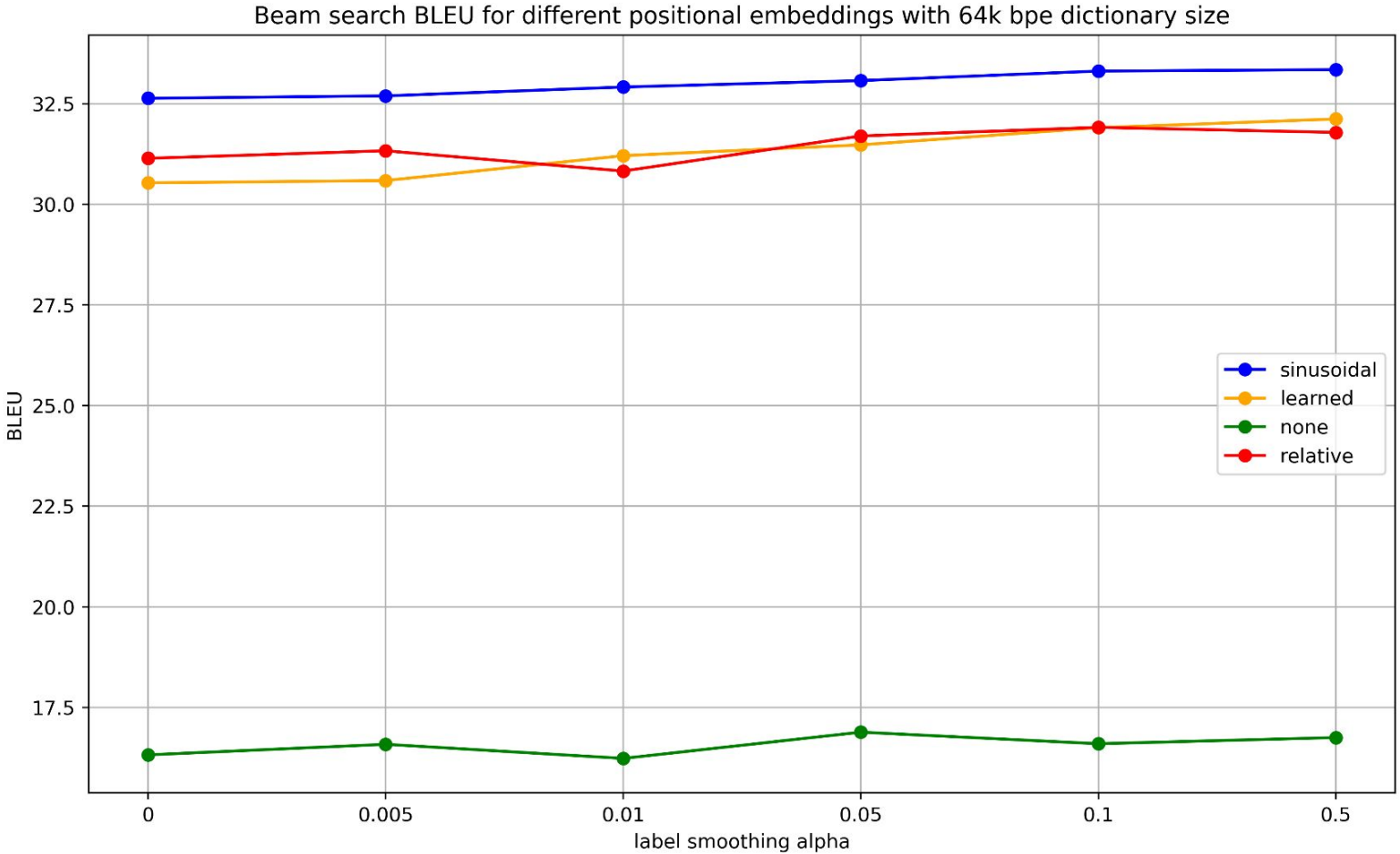

Beam search mean sentence length for different positional embeddings with 4k bpe dictionary size
sinusoid
learned
none
relative
reference
length
26.0
25.5
25.0
24.5
24.0
23.5
23.0
22.5
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha

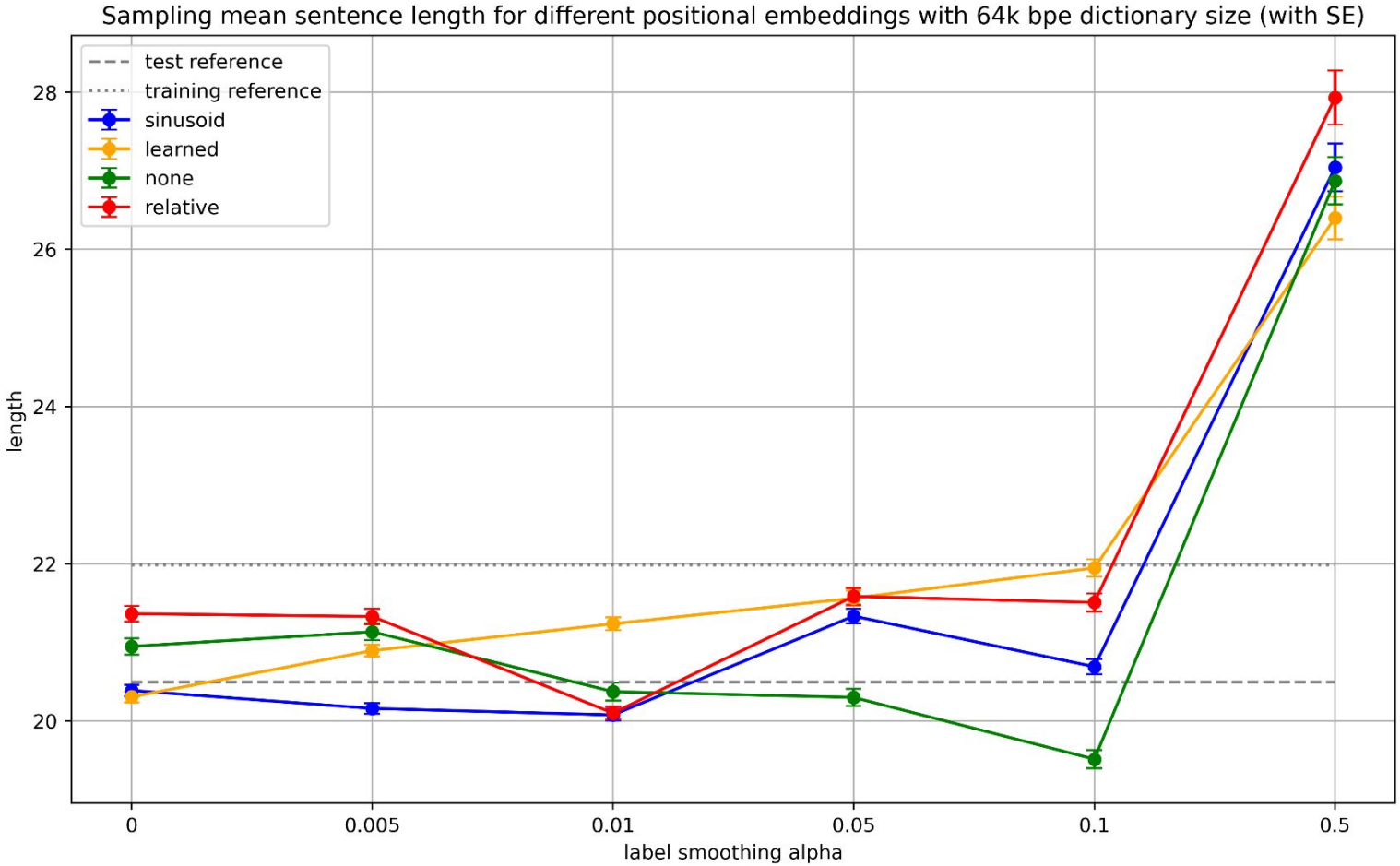

Beam search BLEU for different positional embeddings with 4k bpe dictionary size
BLEU
35.0
32.5
30.0
27.5
25.0
22.5
20.0
17.5
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha
sinusoidal
learned
none
relative

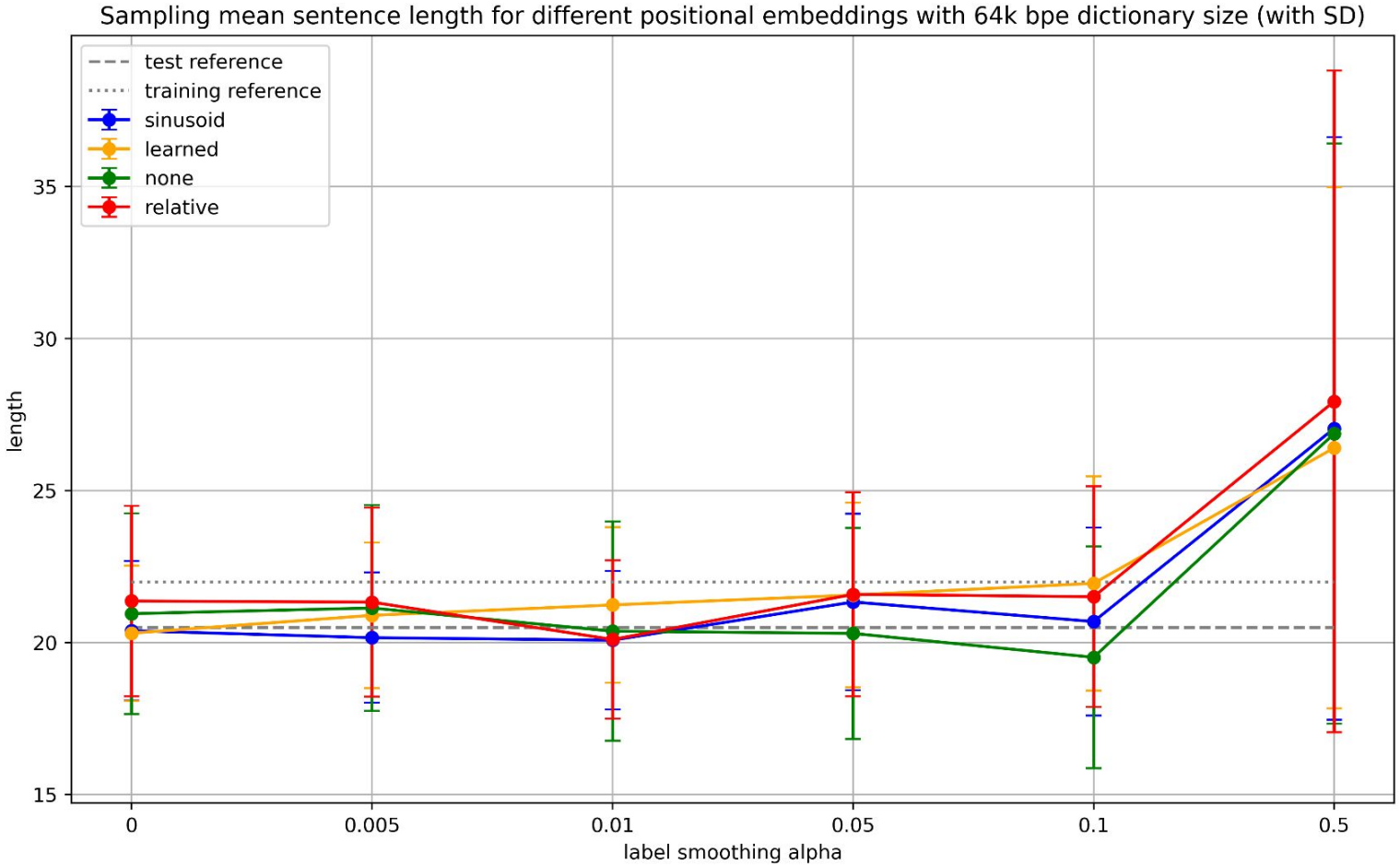

Sampling mean sentence length for different positional embeddings with 4k bpe dictionary size (with SE)
test reference
training reference
sinusoid
learned
none
relative
length
42.5
40.0
37.5
35.0
32.5
30.0
27.5
25.0
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha


Sampling mean sentence length for different positional embeddings with 4k bpe dictionary size (with SD)
test reference
training reference
sinusoid
learned
none
relative
length
label smoothing alpha
0
0.005
0.01
0.05
0.1
0.5
20
25
30
35
40
45
50
55
60


Beam search mean sentence length for different positional embeddings with 64k bpe dictionary size
sinusoid
learned
none
relative
reference
length
20.5
20.0
19.5
19.0
18.5
18.0
17.5
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha


Beam search BLEU for different positional embeddings with 64k bpe dictionary size
BLEU
32.5
30.0
27.5
25.0
22.5
20.0
17.5
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha
sinusoidal
learned
none
relative


Sampling mean sentence length for different positional embeddings with 64k bpe dictionary size (with SE)
test reference
training reference
sinusoid
learned
none
relative
length
28
26
24
22
20
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha


Sampling mean sentence length for different positional embeddings with 64k bpe dictionary size (with SD)
test reference
training reference
sinusoid
learned
none
relative
length
35
30
25
20
15
0
0.005
0.01
0.05
0.1
0.5
label smoothing alpha