ngram baseline

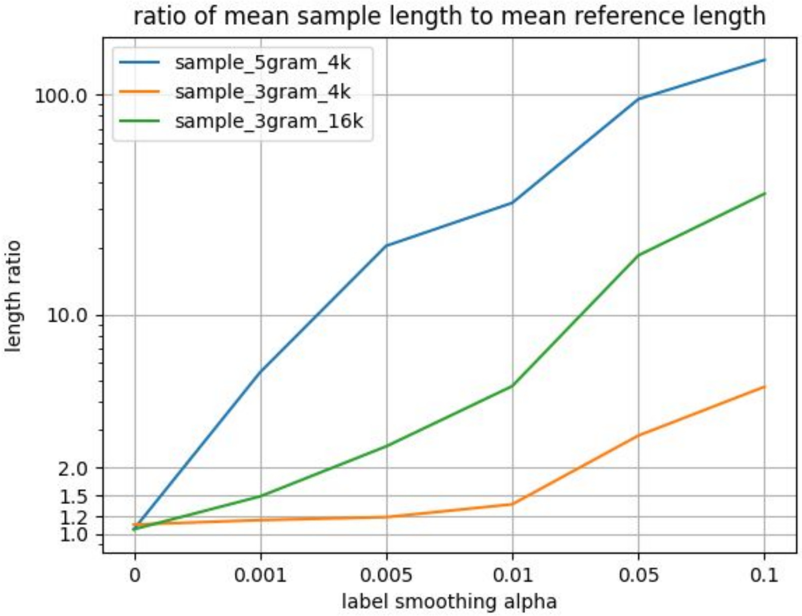
ngram baseline

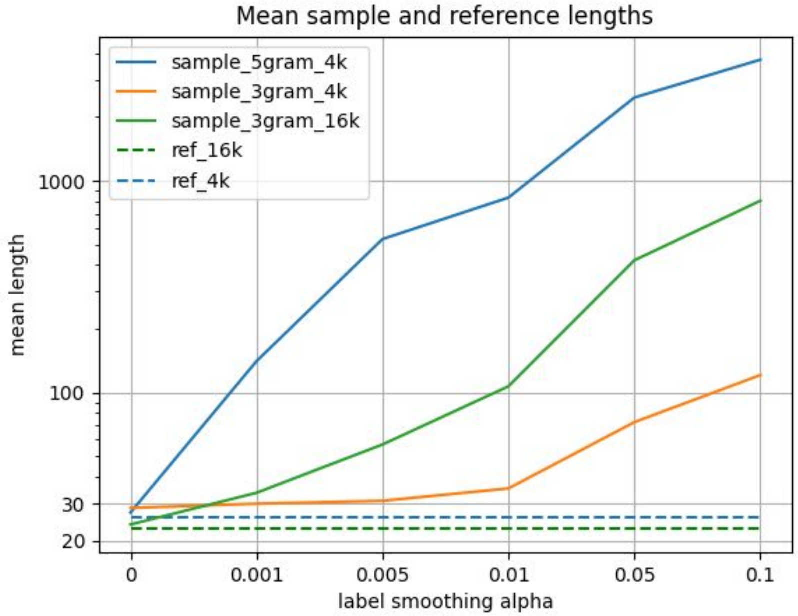
- Ngram model trained on iwslt17 de-en english data (same data as Riley and Chiang models)
- Label smoothing implemented as linear interpolation between model probability and uniform distribution during sampling:

  ```
  log_probs_smoothed = np.log((1-alpha) * np.exp(log_probs) + alpha * (1/root.max_idx))
  ```

- Results:
  - All results based on bpe tokens
  - length ratio: (len(generated)/count(generated)) / (len(reference)/count(reference))
  - mean: (len(generated)/count(generated))
- Different configurations:
  - 3/5-grams
  - vocab size 4k / 16k
  - ls alpha: 0, 0.001, 0.005, 0.01, 0.05, 0.1

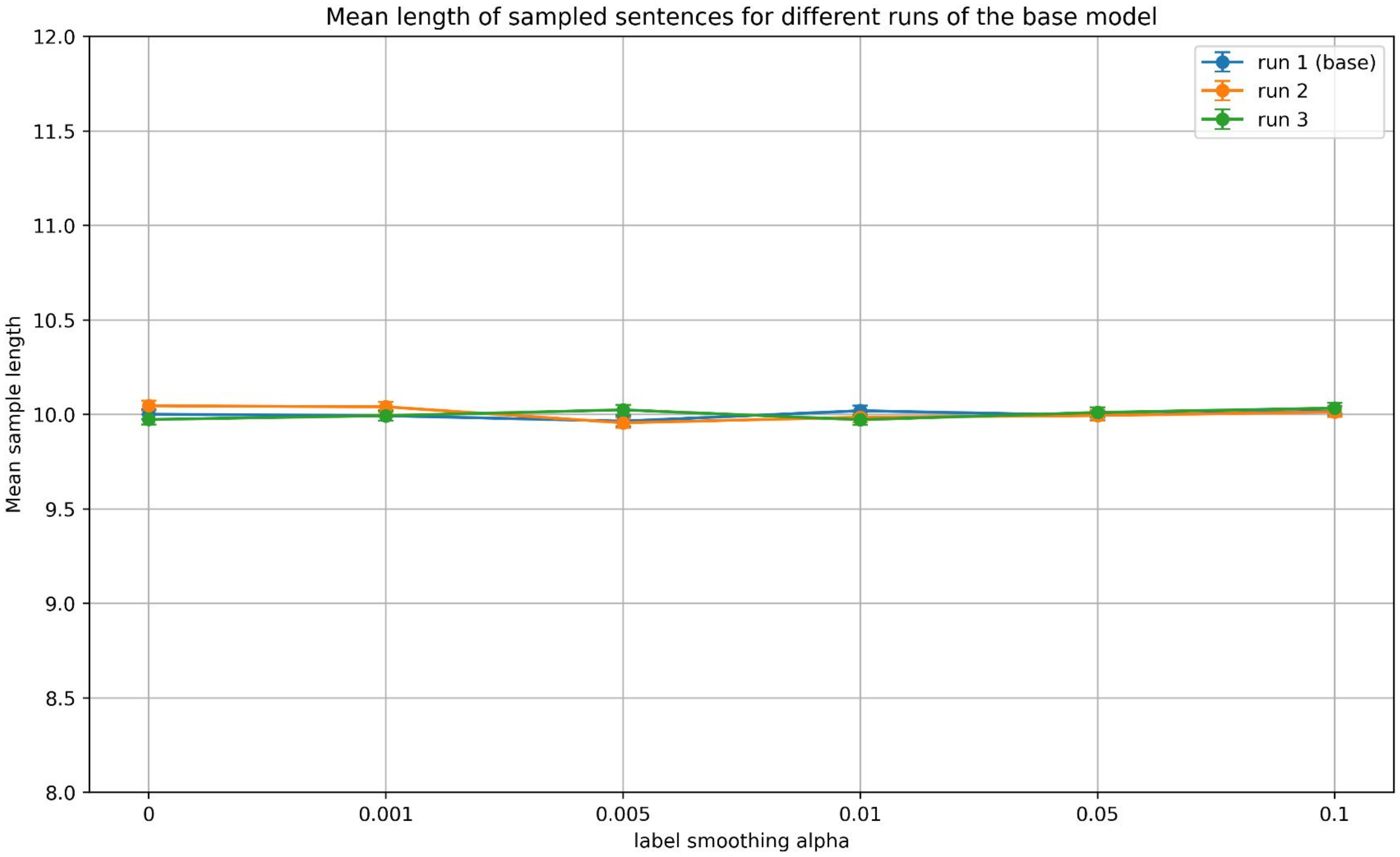
length ratio and mean

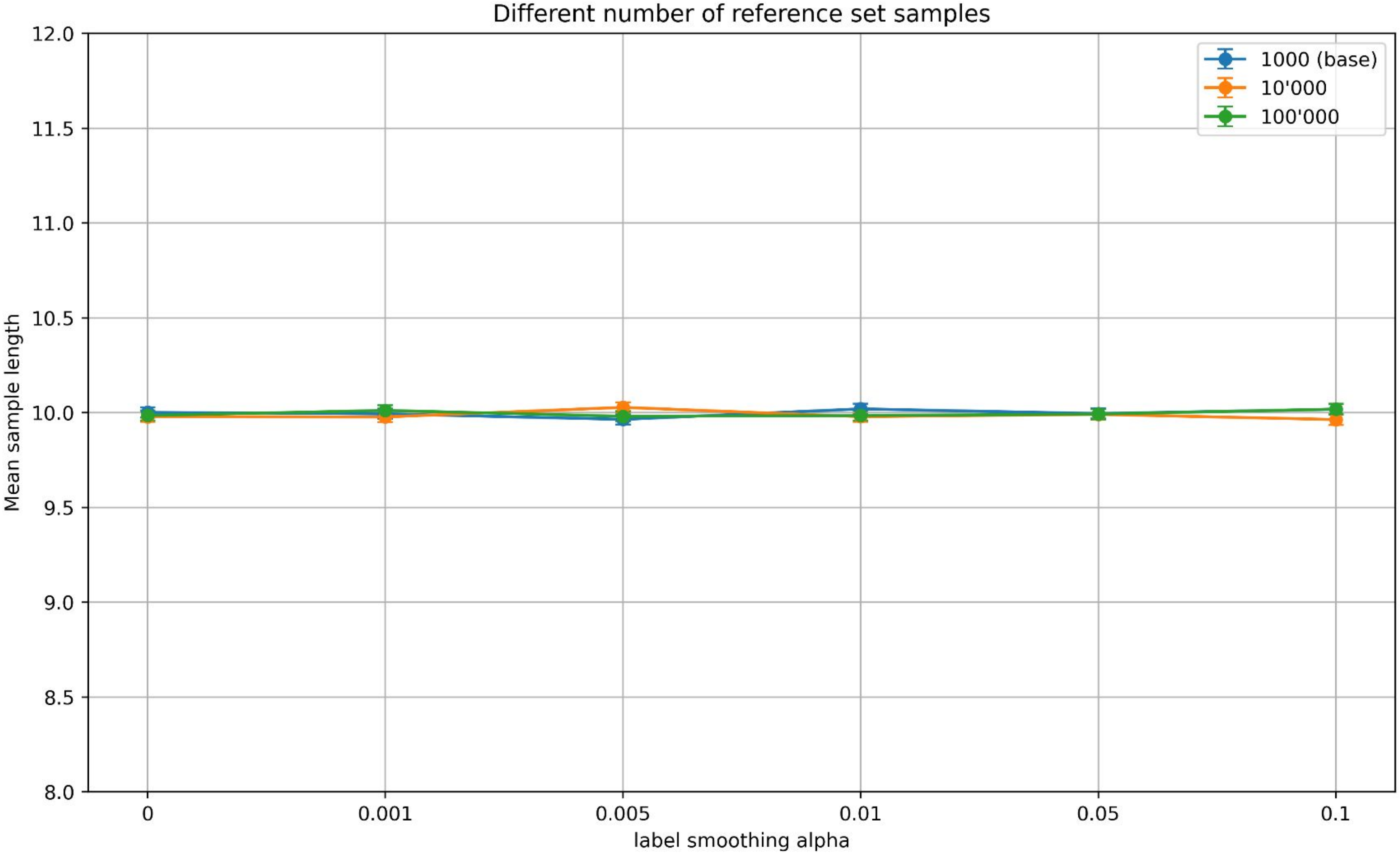

ratio of mean sample length to mean reference length

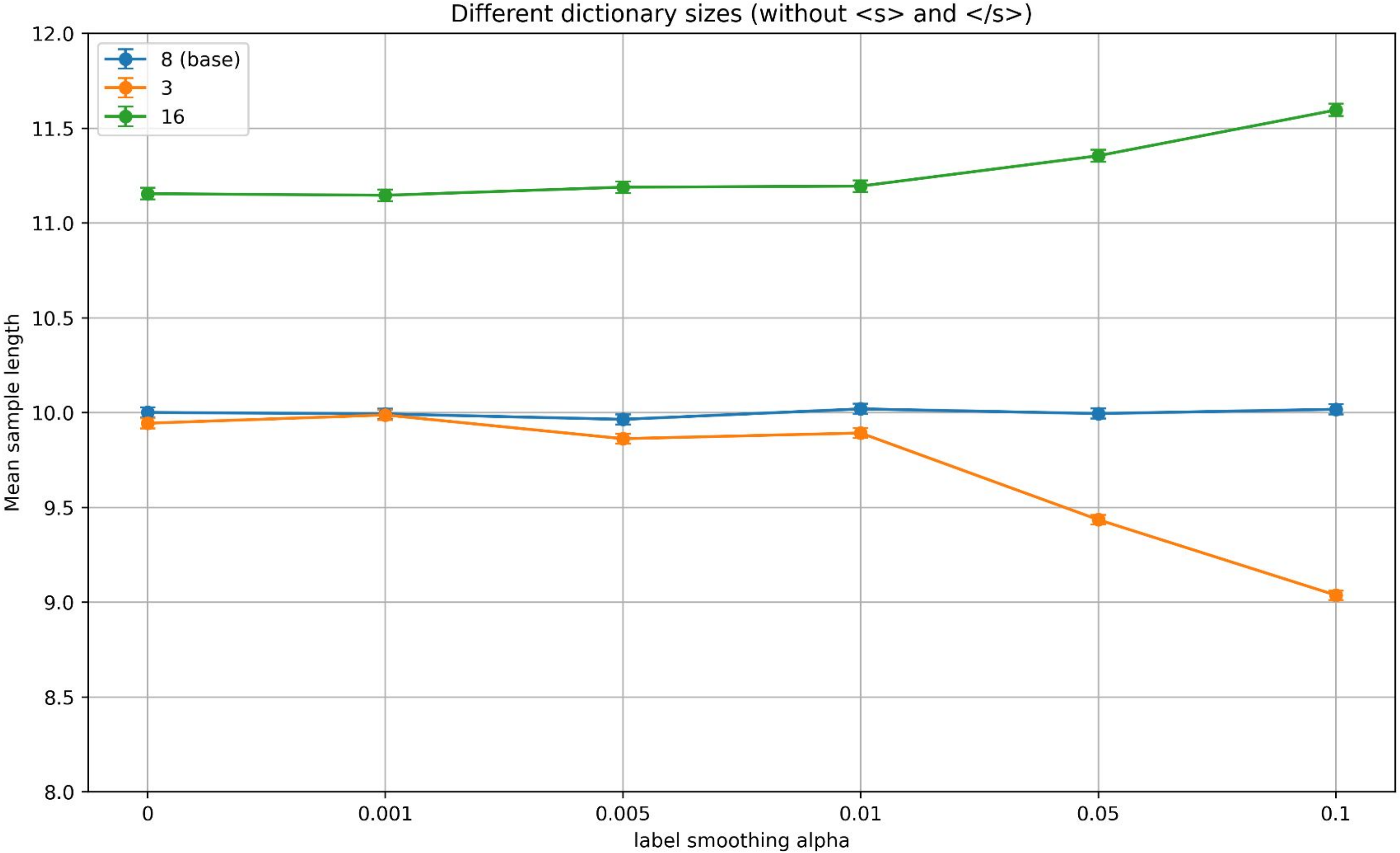

Mean sample and reference lengths

Artificial data ngram analysis

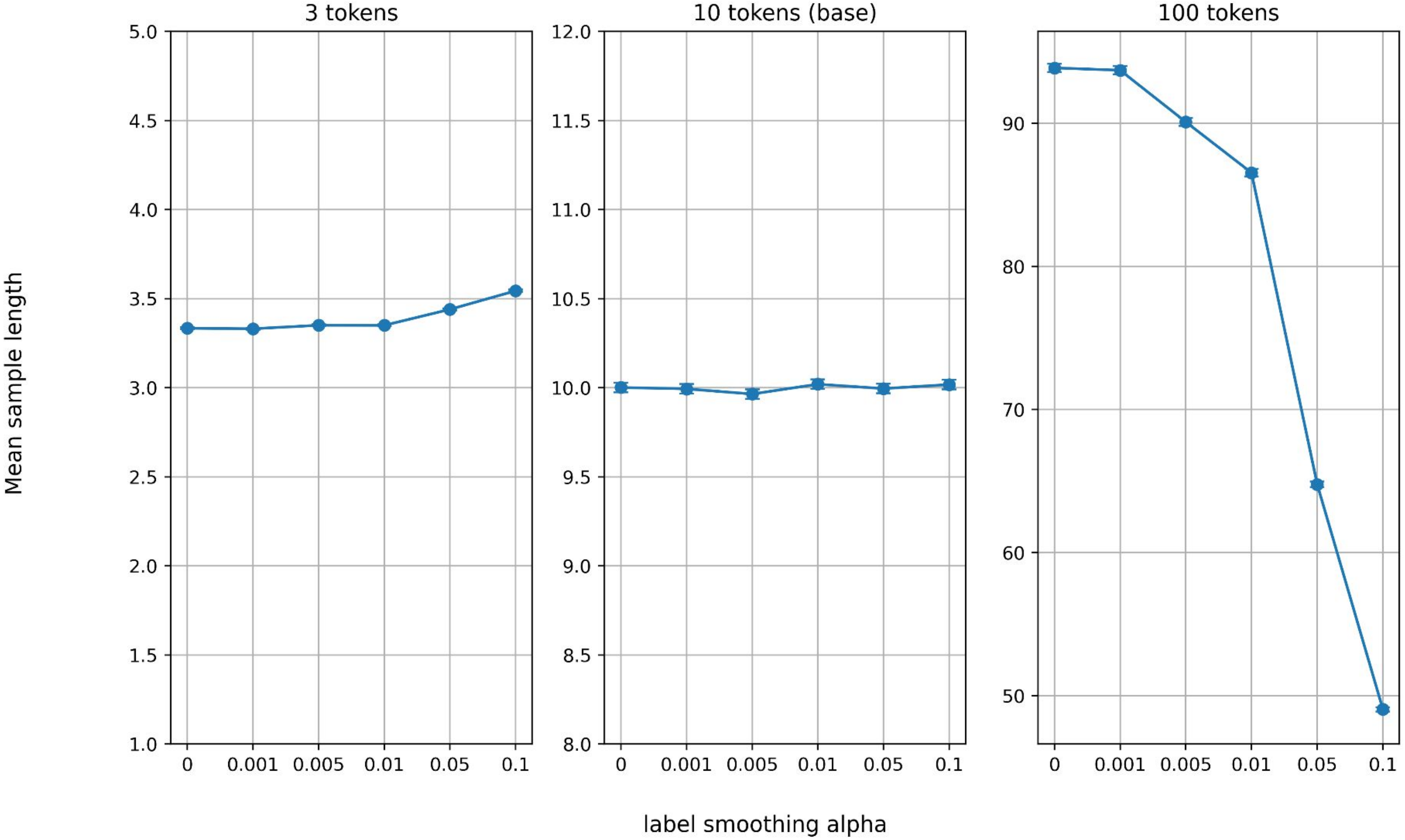
- Artificial dataset consisting of samples of equal length
- Base setup: 8 dictionary tokens, 10 tokens per reference length, 1000 samples in the dataset. 3gram trained with different label smoothing alpha settings.
- We sample 1000 sentences from each model and report the mean length of sampled sentences and the estimated standard error of the sample mean
- We vary several aspects and compare them:
  - Variance across runs
  - Number of sentences in the reference set
  - Number of dictionary tokens
  - Ngram history size n
  - Reference sentence lengths


Mean length of sampled sentences for different runs of the base model
Mean sample length
12.0
11.5
11.0
10.5
10.0
9.5
9.0
8.5
8.0
0
0.001
0.005
0.01
0.05
0.1
label smoothing alpha
run 1 (base)
run 2
run 3


Different number of reference set samples
Mean sample length
12.0
11.5
11.0
10.5
10.0
9.5
9.0
8.5
8.0
label smoothing alpha
0
0.001
0.005
0.01
0.05
0.1
1000 (base)
10'000
100'000


Different dictionary sizes (without <s> and </s>)
8 (base)
3
16
Mean sample length
12.0
11.5
11.0
10.5
10.0
9.5
9.0
8.5
8.0
label smoothing alpha
0
0.001
0.005
0.01
0.05
0.1

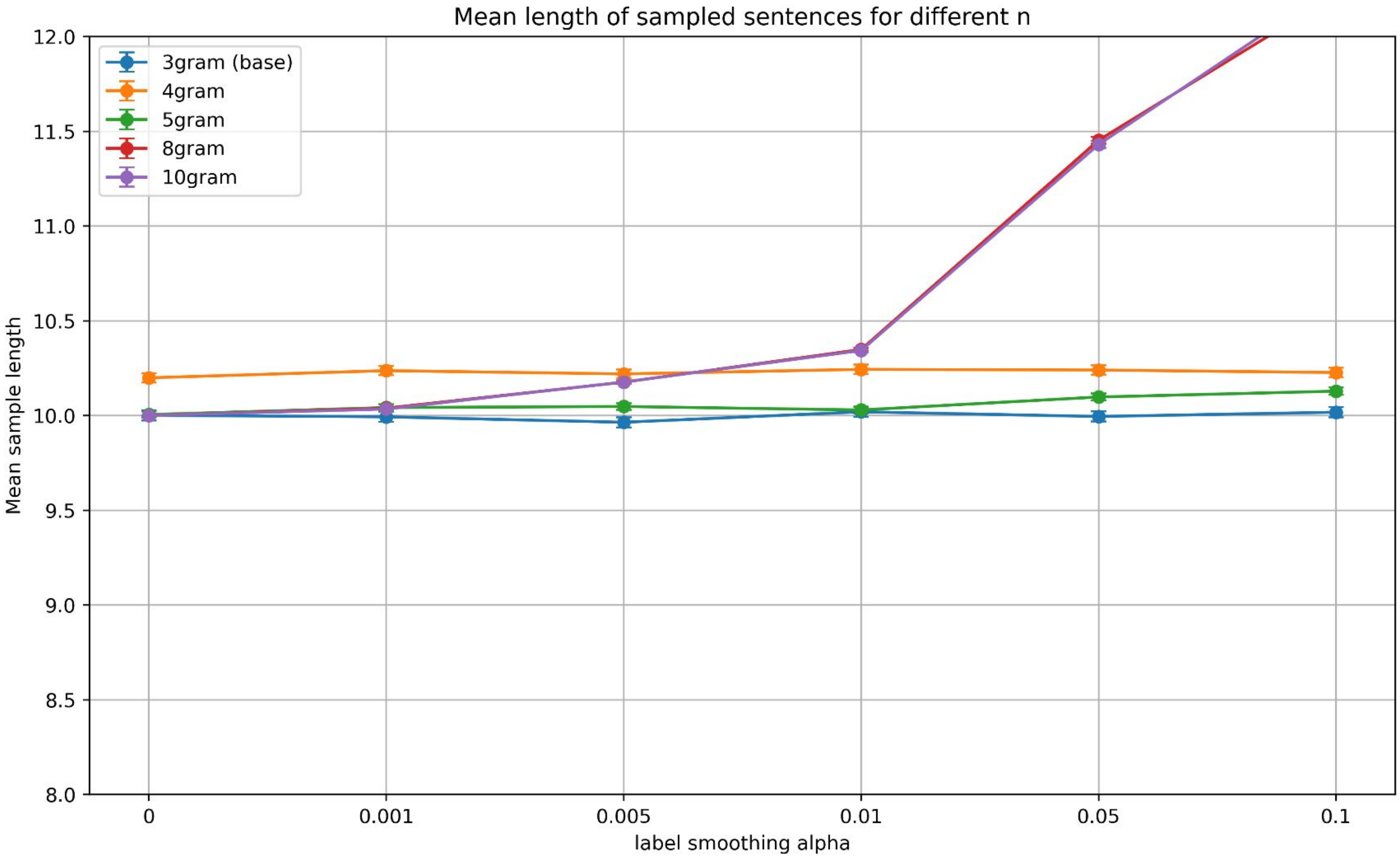
Differing number of reference bpe tokens per sample



3 tokens

10 tokens (base)

100 tokens


Mean length of sampled sentences for different n
3gram (base)
4gram
5gram
8gram
10gram
Mean sample length
12.0
11.5
11.0
10.5
10.0
9.5
9.0
8.5
8.0
label smoothing alpha
0
0.001
0.005
0.01
0.05
0.1