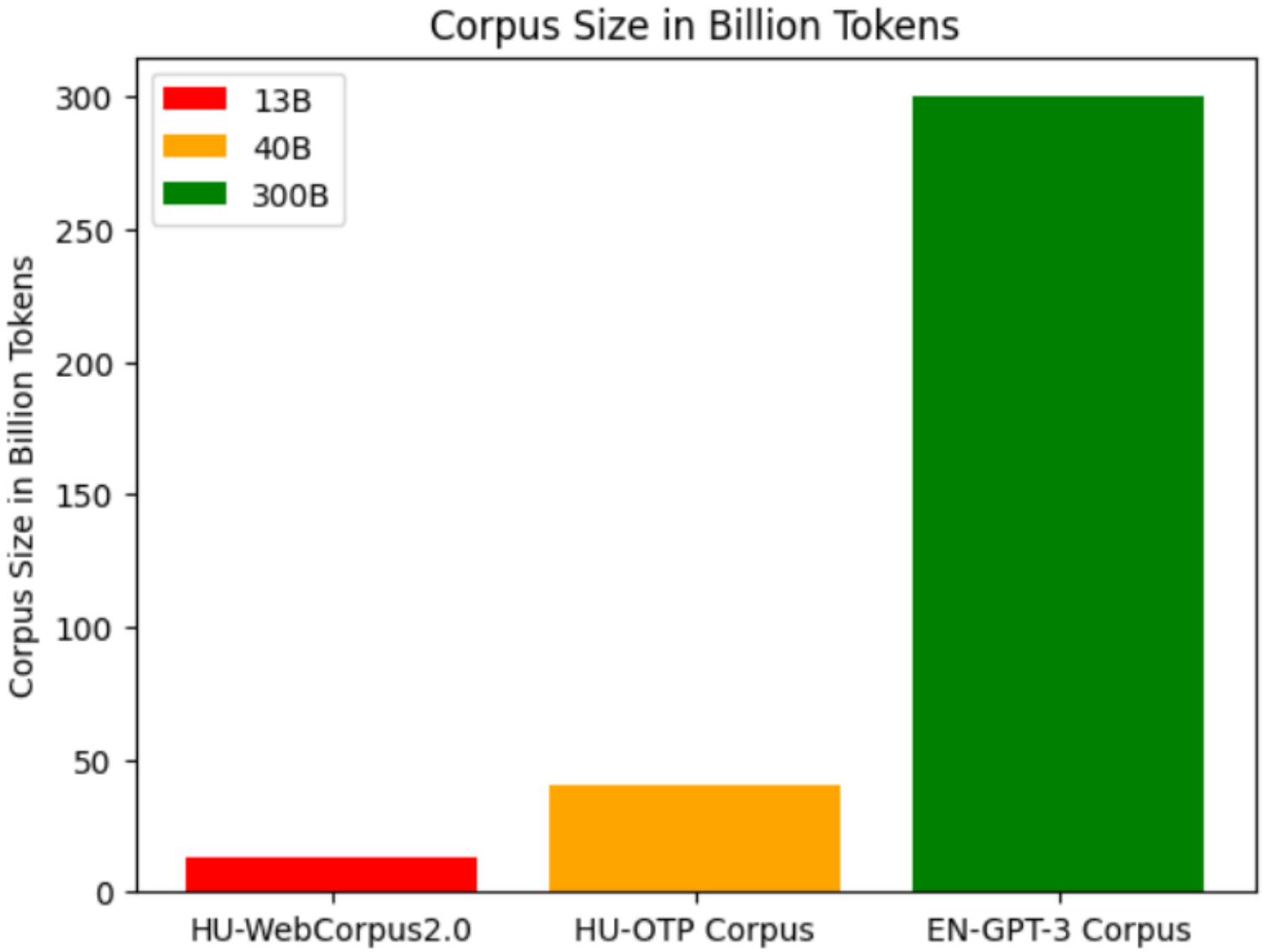
OTP Bank

Schin Lotár
AI CoE Lead

Corpus


Corpus Size in Billion Tokens
13B
40B
300B
Corpus Size in Billion Tokens
300
250
200
150
100
50
0
HU-WebCorpus2.0
HU-OTP Corpus
EN-GPT-3 Corpus

Corpus size vs sequence size

- 1.5B model’s sequence size is 1024M
- 13B model’s sequence size is 2048M

- WebCorpus 2.0 has 13B tokens

-> if seq size = 1M -> **12,695,312 sequences**

-> if seq size = 2M -> **6,347,656 sequences**

Sequence size vs epochs

otpbank

Batch size is set to: 1024

1 epoch = (# of sequences / batch size) steps

Seq. size: 1M -> 1 epoch = 12,695,312 / 1024 = **12 397**
Seq. size: 2M -> 1 epoch = 6,347,656 / 1024 = **6 198**

How much tokens do we need?

Sambanova’s experience is:

- 13B model (300k steps) can outperform 175B model (150k steps), but…

300k steps -> 1024*300k*2048=**629B tokens!!!**

150k steps -> **314B tokens!!!**

What about context window?

- How much tokens can be looked at once is important
- The bigger the context window the more compute power needed
- GPT-4 with 32,768 context size can check ~50 A4 pages!

What are our next steps?

- Moving towards multilingual models
- Checking larger context windows
- Taking advantage of larger Hungarian Corpus

Thanks for your attention!

Q&A

Schin Lotár
AI CoE Lead