SKAI LAB

FLEXITOKENS: Flexible Tokenization for Evolving Language Models

Abraham Toluwase Owodunni[♠], **Orevaoghene Ahia**[♣], **Sachin Kumar**[♠]
The Ohio State University[♠], University of Washington[♣]

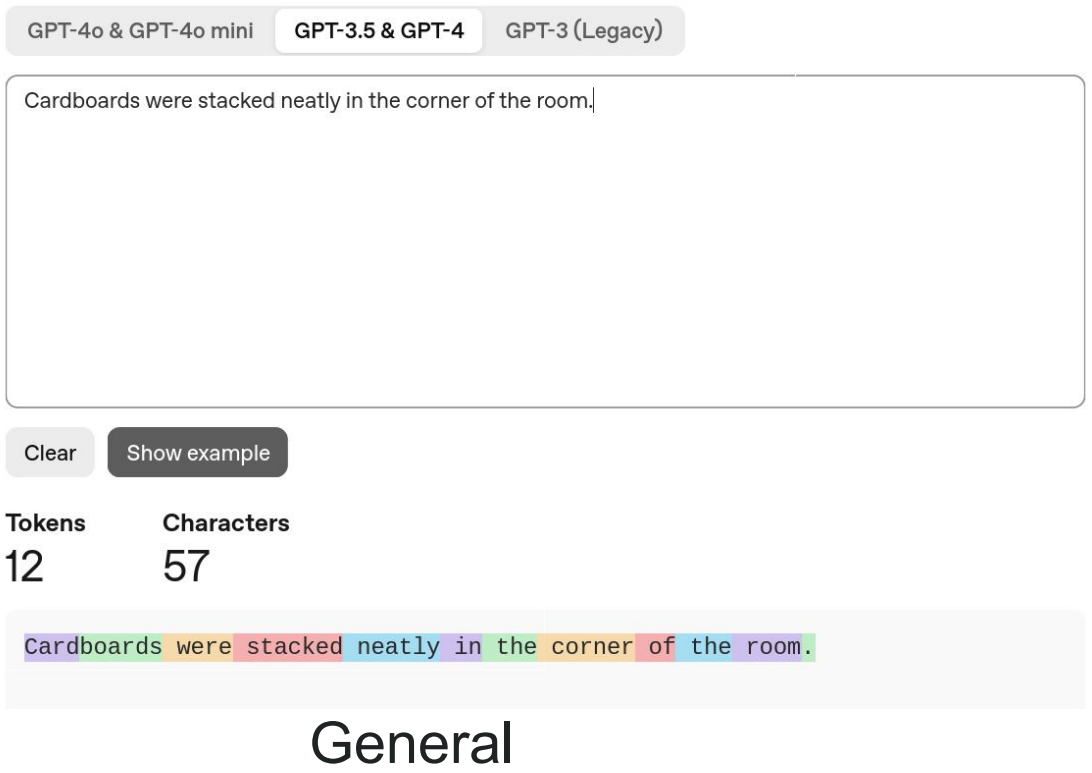
Friday, November 19th, 2025

November 2025

https://arxiv.org/abs/2507.12720

Tokenization & Domains

- **Most modern LMs use a Byte-level BPE, they oversegment non-general domains**



General

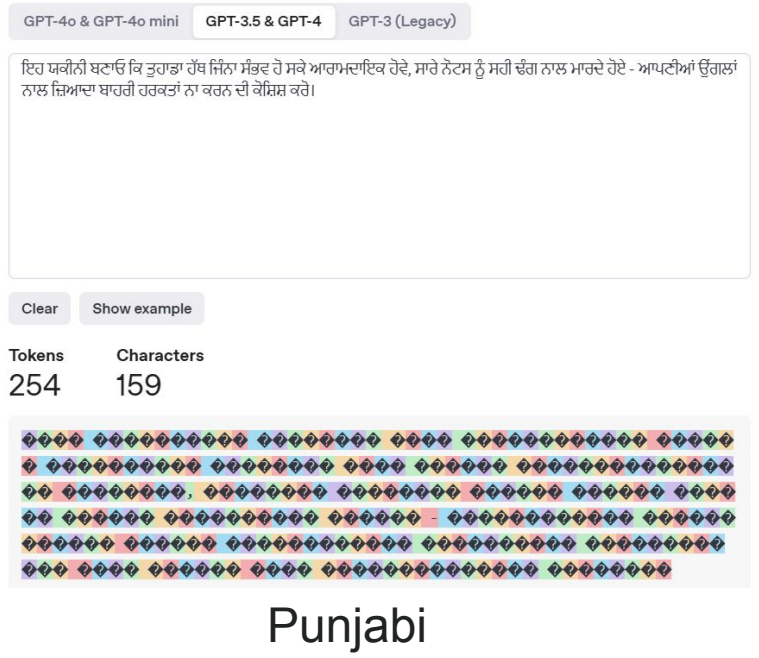
Health

Tokenization & Languages

- **Byte-level BPE laos makes vocabulary construction language agnostic.**

English

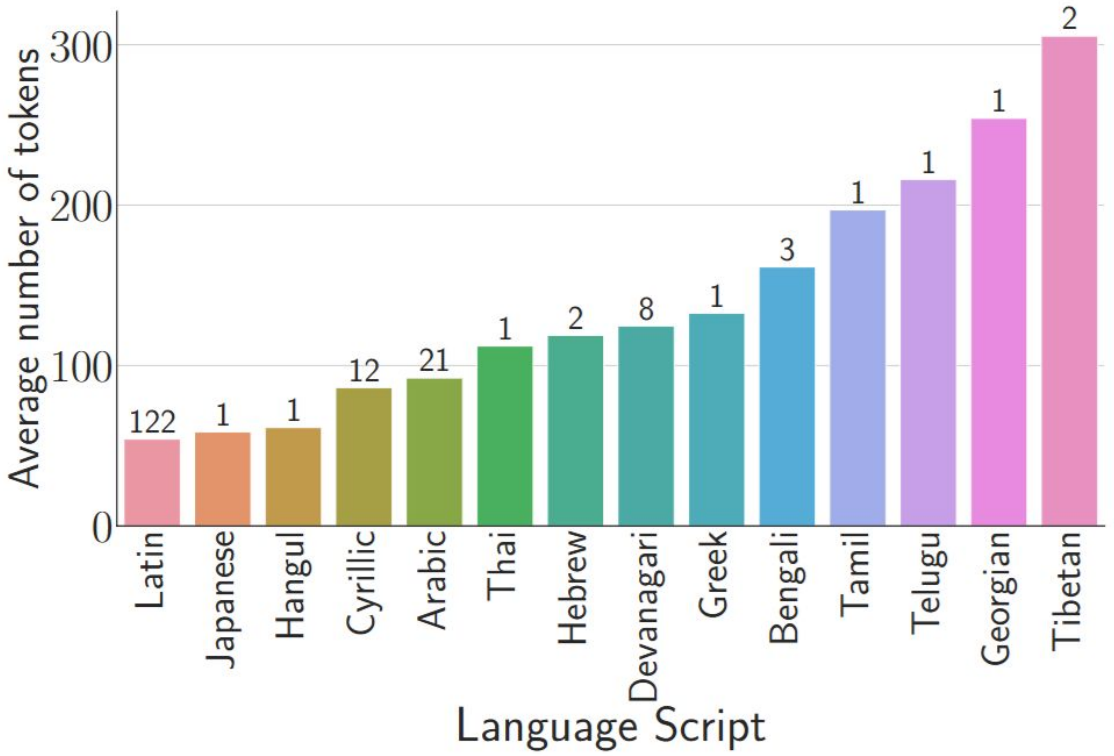
Japanese



Punjabi

Tokenization & Languages

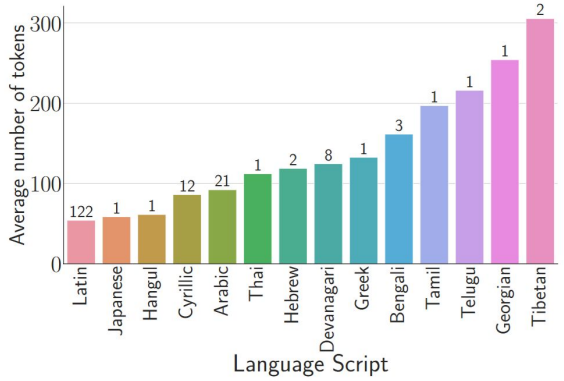
- **Sequence lengths across languages**



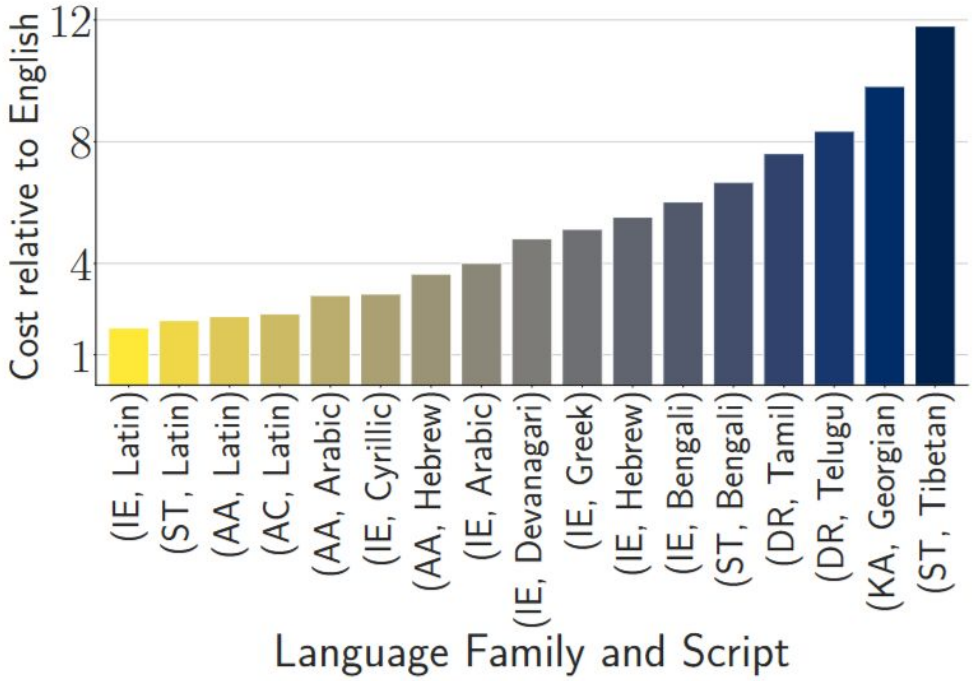
Ahia, Orevaoghene, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. "Do all languages cost the same? tokenization in the era of commercial language models." EMNLP 2023

Tokenization & Languages

- **Sequence lengths across languages**
- **Cost shoots up**



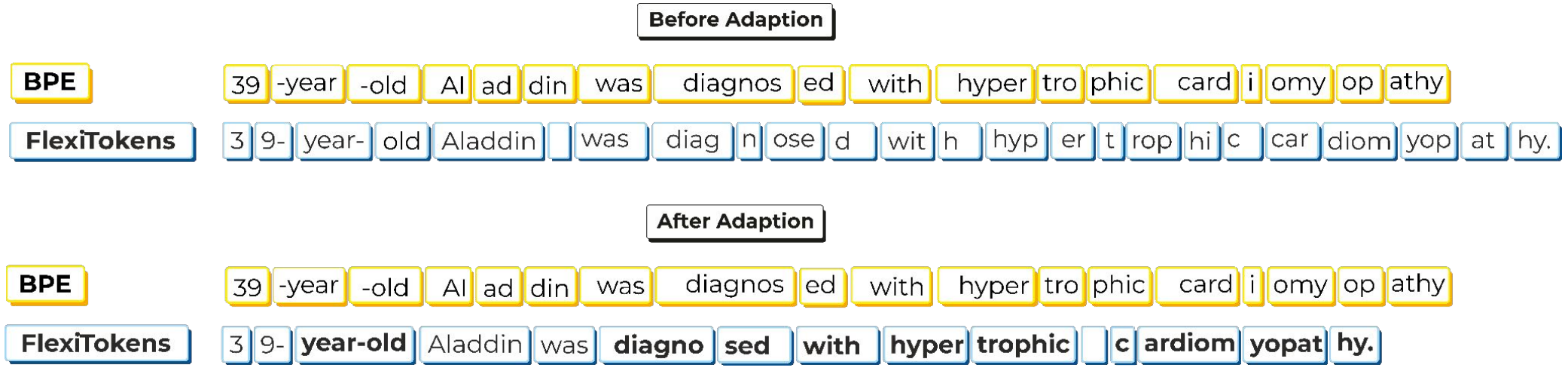
LLM APIs charge per token



Ahia, Orevaoghene, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. "Do all languages cost the same? tokenization in the era of commercial language models." EMNLP 2023

We want Flexible Tokenization

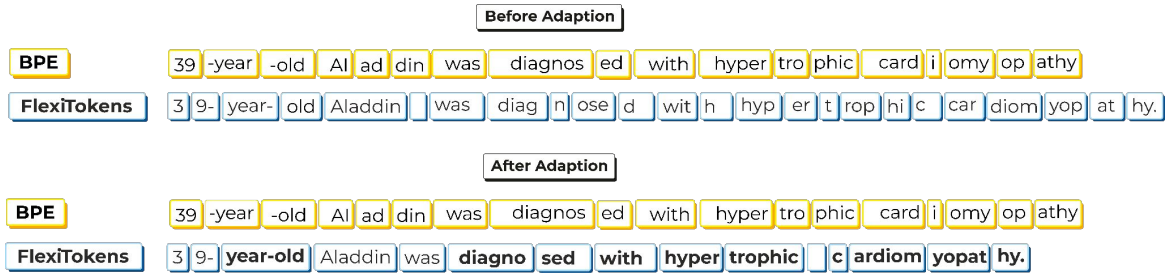
- **A model that adapts its tokenization to the target domain (language, medical) & capture more meaningful context effectively**
- **Produces less fragmented sequence of tokens than BPE**



We want Flexible Tokenization

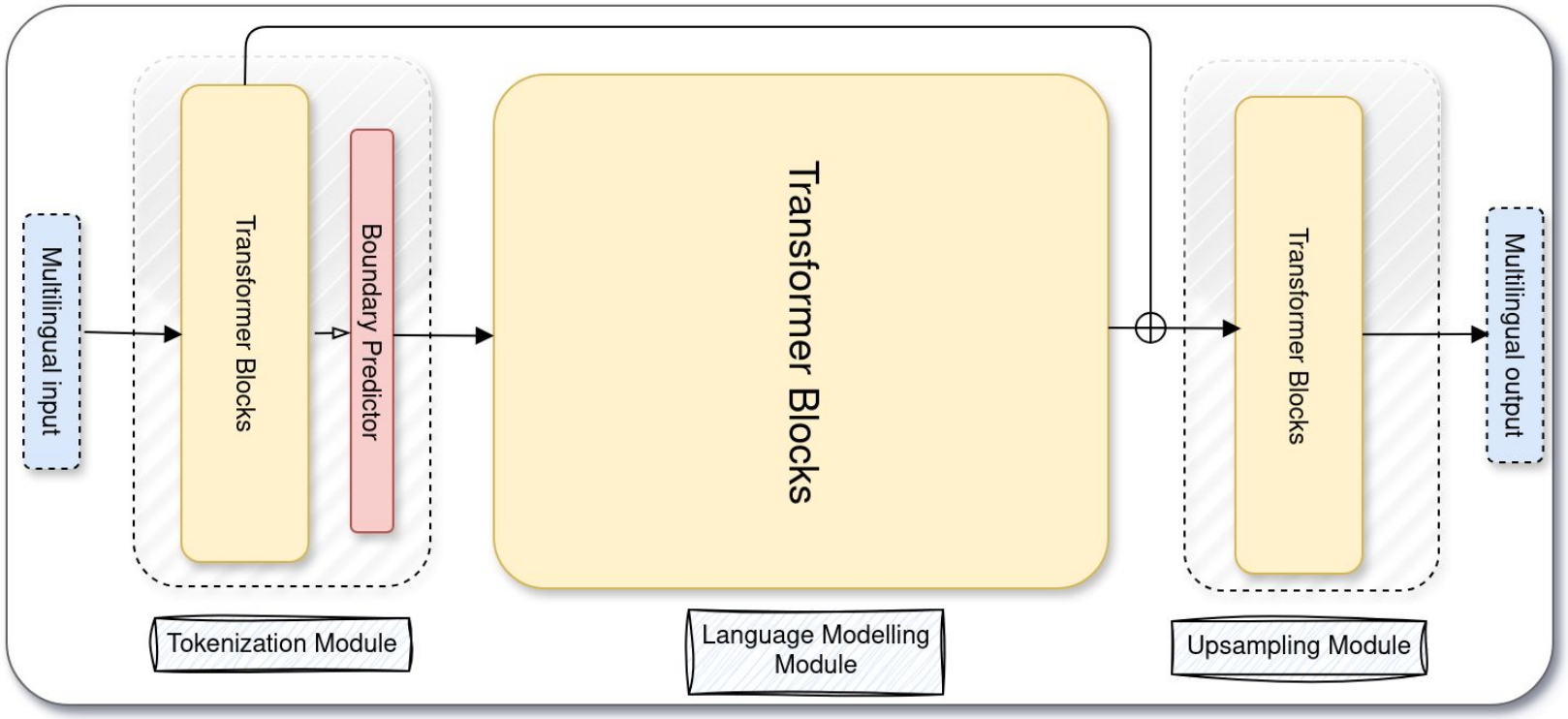
- A model that adapts its tokenization to the target domain (language, medical) & capture more meaningful context effectively
- Produces less fragmented sequence of tokens than BPE

- **Tokenizer expansion is leads to forgetting, Non-latin languages always suffer.**



How?

- Train a model that allows flexible tokenization
- We use an hierarchical byte-level model with gradient based tokenisation.
-

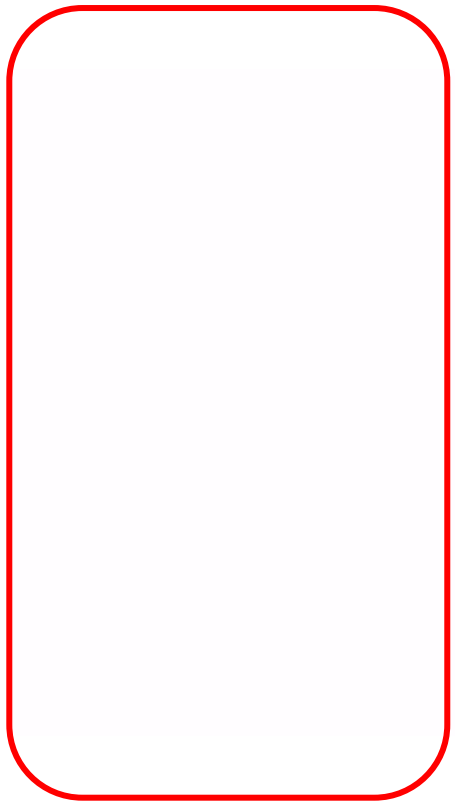


How?

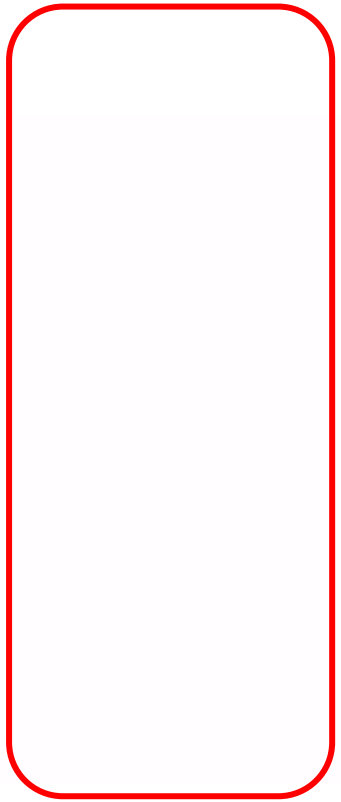
- **Specifically, we use the Hourglass hierarchical transformer model.**

Piotr Nawrot et al.. "Efficient Transformers with Dynamic Token Pooling"ACL 2023

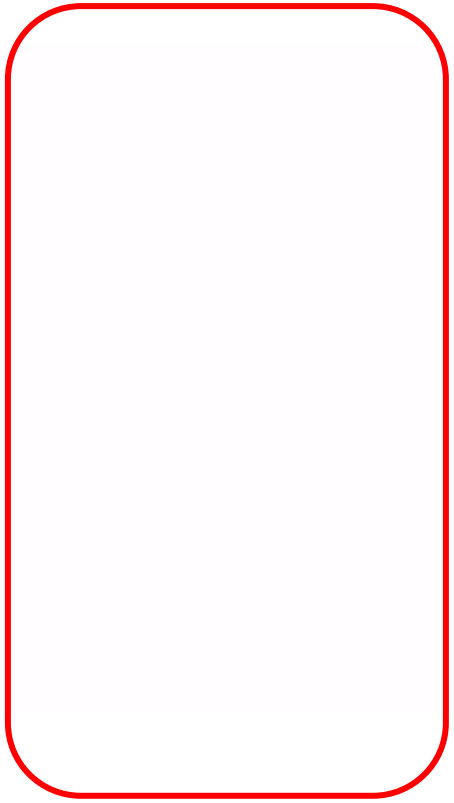
- Tokenization submodule
-



- Tokenization submodule
- Language Modeling submodule
-



- Tokenization submodule
- Language Modeling submodule
- Upsampling submodule



Model Training : Controlling Tokenization

- NTP Loss + BP Loss
  - Previous works: Magnet (Ahia et al, 204)

$$\sum_{i=1}^{N} -\log p_\theta(x_i|x_{<i}) - \lambda \sum_{S} \mathbb{I}(\text{script}(\mathbf{x}) = S) \log \text{Binomial}(\beta_S; N, k)$$

NTP Loss

BP Loss

$$\text{Binomial}(\beta; N, k) = \binom{N}{k} \beta^k (1-\beta)^{N-k}, \quad \text{and} \quad k = \sum_{N} b_t.$$

Where:

– λ is the weighting coefficient for the regularization term

– S is the script or language group (e.g., Latin, Cyrillic, Arabic)

– $\mathbf{I}(\backslash\text{script}(x) = S)$ is the indicator function equal to 1 if the input uses script S

– $\text{Binomial}(\beta_S; N, k)$ is the Binomial distribution giving the probability of k boundaries out of N tria

with parameter β_S

– β_S is the script-specific probability of a segment boundary

– k is the number of segment boundaries predicted in the sequence

Model Training

$$p(b_t = 1) = \text{sigmoid}\left(\text{MLP}\left(\boldsymbol{h}_t\right)\right)$$

Boundary Predictor makes discrete non-differentiable decisions.

Make the network differentiable using the “gumbel softmax trick”

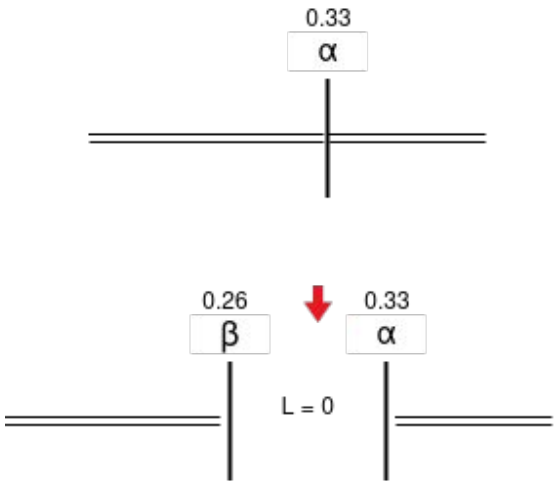
Boundary predictor includes a hyperparameter α that helps control the segmentation rate: α = 1 (byte level); α = 0 (the whole sequence is one token)

Model Training across Language Scripts

- NTP Loss + BP Loss
  - Previous works (Limitations):
    - Inflexible tokenization,
    - How many BPs do we need for 20 script?

Model Training

- NTP Loss + BP Loss
  - Our work (FlexiTokens):
  - **Simpler Loss**



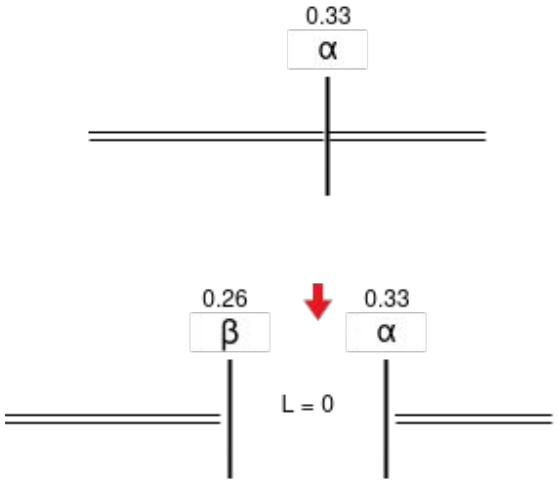
Model Training

- NTP Loss + BP Loss
  - Our work (FlexiTokens):
  - **Simpler Loss**

$$\sum_{i=1}^{N} -\log p_\theta(x_i|x_{<i}) - \lambda \sum_{S} \mathbb{I}(\text{script}(\mathbf{x}) = S) \log \text{Binomial}(\beta_S; N, k)$$

$$\mathcal{L} = \sum_{i=1}^{N} -\log p_\theta(x_i \mid x_{<i}) + \sum_{\mathcal{M}} \mathbb{I}(\text{language}(\mathbf{x}) = L)\mathcal{L}_{\mathcal{BP}\,L}$$

$$\mathcal{L}_{\mathcal{BP}} = \max\left(\frac{k}{N} - \alpha, 0\right) + \max\left(\beta - \frac{k}{N}, 0\right), \text{ where } \beta = \alpha - \lambda\sigma$$



Model Training

- NTP Loss + BP Loss
  - NTP Loss + BP Loss
  - Our work (FlexiTokens):

$$\mathcal{L} = \sum_{i=1}^{N} -\log p_\theta(x_i \mid x_{<i}) + \sum_{\mathcal{M}} \mathbb{I}(\text{language}(\mathbf{x}) = L)\mathcal{L}_{\mathcal{BP}\,L}$$

$$\mathcal{L}_{\mathcal{BP}} = \max\left(\frac{k}{N} - \alpha, 0\right) + \max\left(\beta - \frac{k}{N}, 0\right), \text{ where } \beta = \alpha - \lambda\sigma$$

– L is the language or dataset group

– $\mathbf{I}(\text{lang}(x) = L)$ is the indicator function for language L

– k/N is the observed boundary (segment) rate for the sample

– α is the target upper-bound compression rate

– β defines the lower-bound ($\beta = \alpha - \lambda\sigma$)

– λ is a hyperparameter controlling penalty slack

– σ is the standard deviation of tokenization rates across samples

Model Training

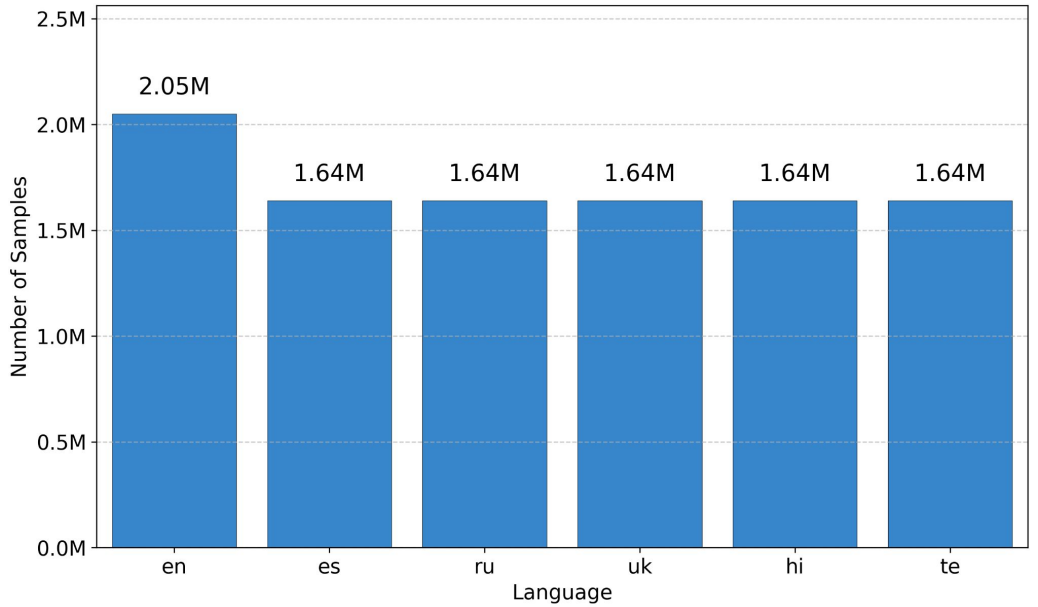
- Computing alpha α and sigma σ
  - Average byte counts ratio of English to other languages in the FLORES (//) dataset.
  - σ is STD of bytes across FLORES dataset.

Table 1: α_L and σ_L values for each language in our training dataset, computed using FLORES-200. The upper bound β_L in Equation 3 is computed as $\alpha_L - \lambda\sigma_L$)

Configuration	en	es	ru	uk	hi	te
FLEXITOKENS 10×	0.1 / 10	0.08 / 12.12	0.05 / 19.92	0.053 / 18.70	0.039 / 25.62	0.037 / 26.91
FLEXITOKENS 5×	0.2 / 5	0.17 / 6.06	0.1 / 9.96	0.107 / 9.35	0.078 / 12.81	0.074 / 13.45
FLEXITOKENS 3×	0.333 / 3	0.28 / 3.64	0.167 / 5.98	0.178 / 5.61	0.13 / 7.68	0.124 / 8.07
σ	0.023	0.019	0.011	0.012	0.009	0.008

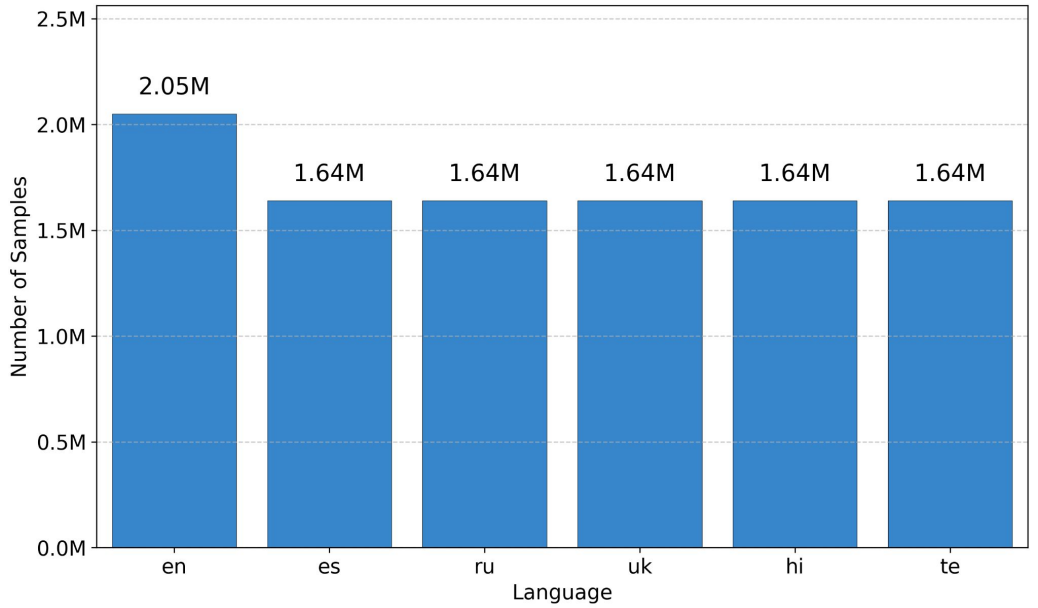
Experiments

- **Data**
  - **26.2 Billion tokens**
  - **Languages:**
    - English, Spanish [Latin]
    - Russian, Ukrainian [Cyrillic]
    - Hindi Telugu [Devanagari, Telugu lipi]
  - **Source: Fineweb & Fineweb2**



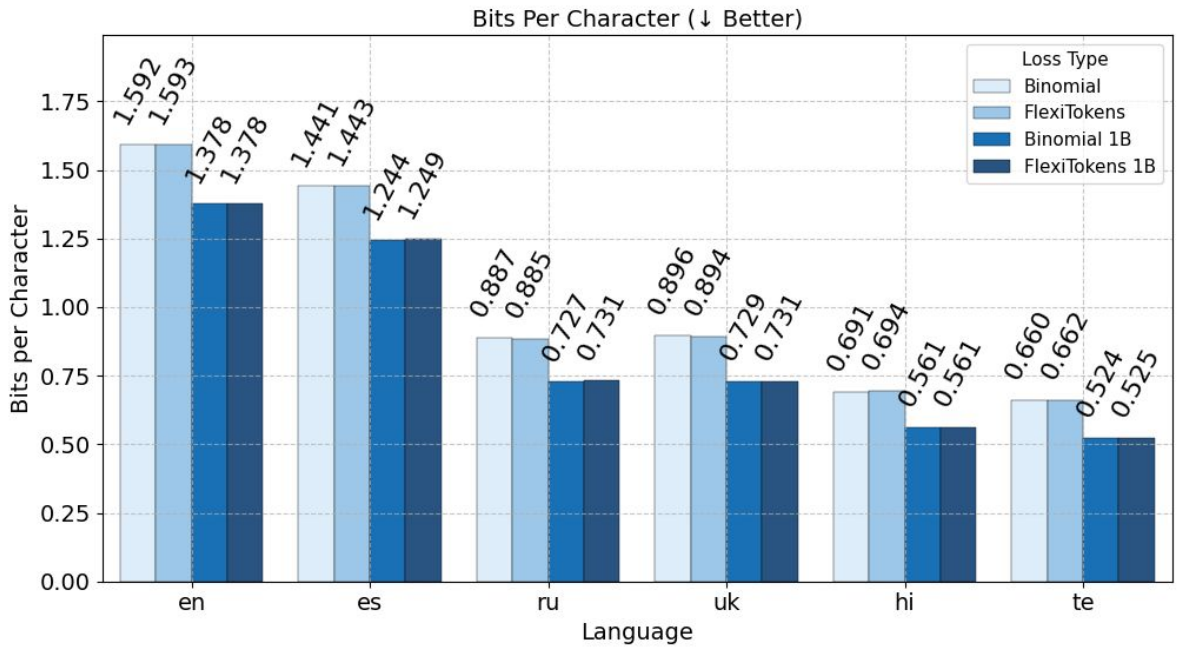
Experiments

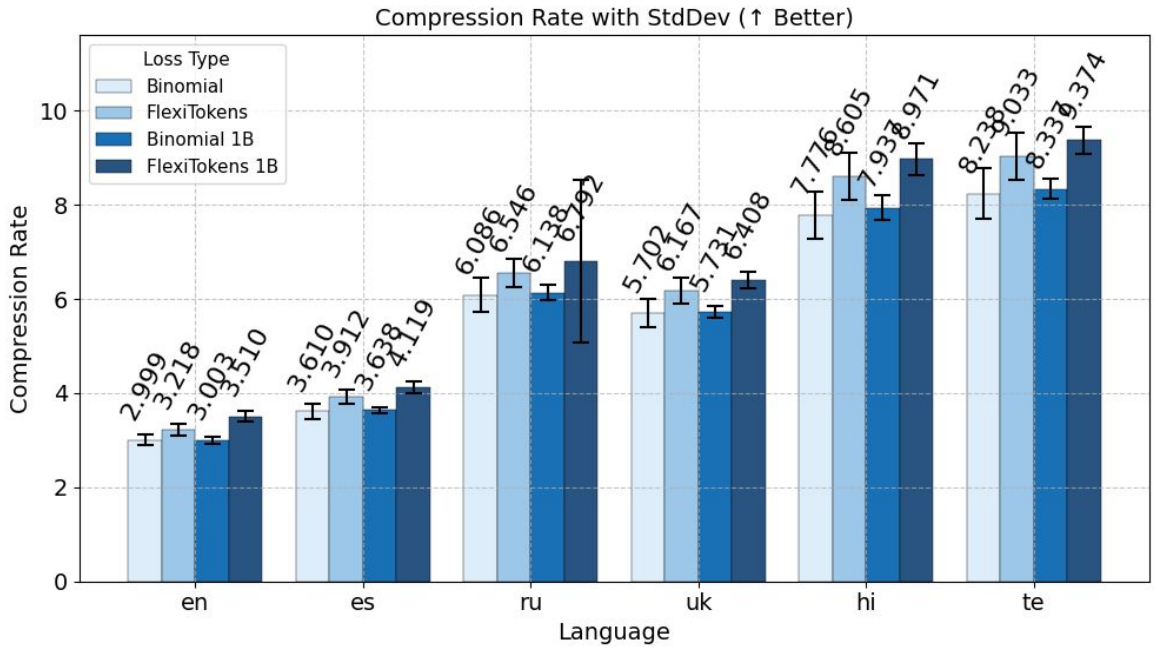
- **Data**
  - **26.2 Billion tokens**
  - **Languages:**
    - English, Spanish [Latin]
    - Russian, Ukrainian [Cyrillic]
    - Hindi Telugu [Devanagari, Telugu lipi]
  - **Source: Fineweb & Fineweb2**
  - **Model:**
    - **116M (1B) params:**
      - Tokenization module: 2 (2)
      - LM module: 12 (24)
      - Upsampling module: 2 (2)
  -



Pretraining Results:

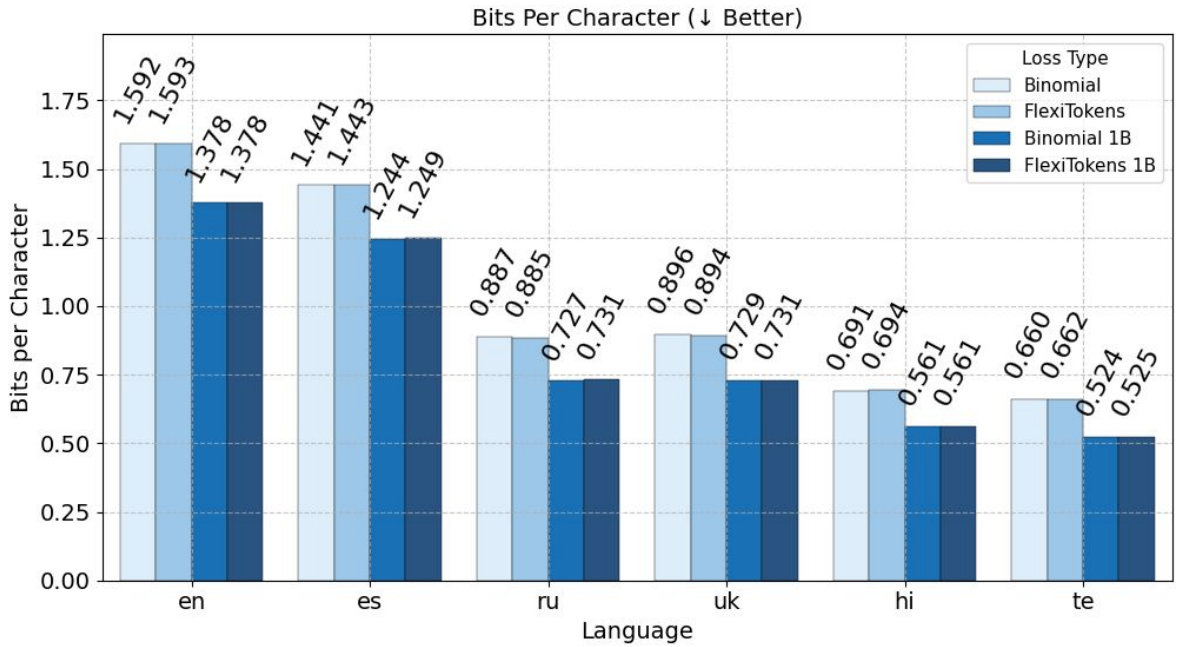
- **Achieve comparable BPC with even higher compression rate**

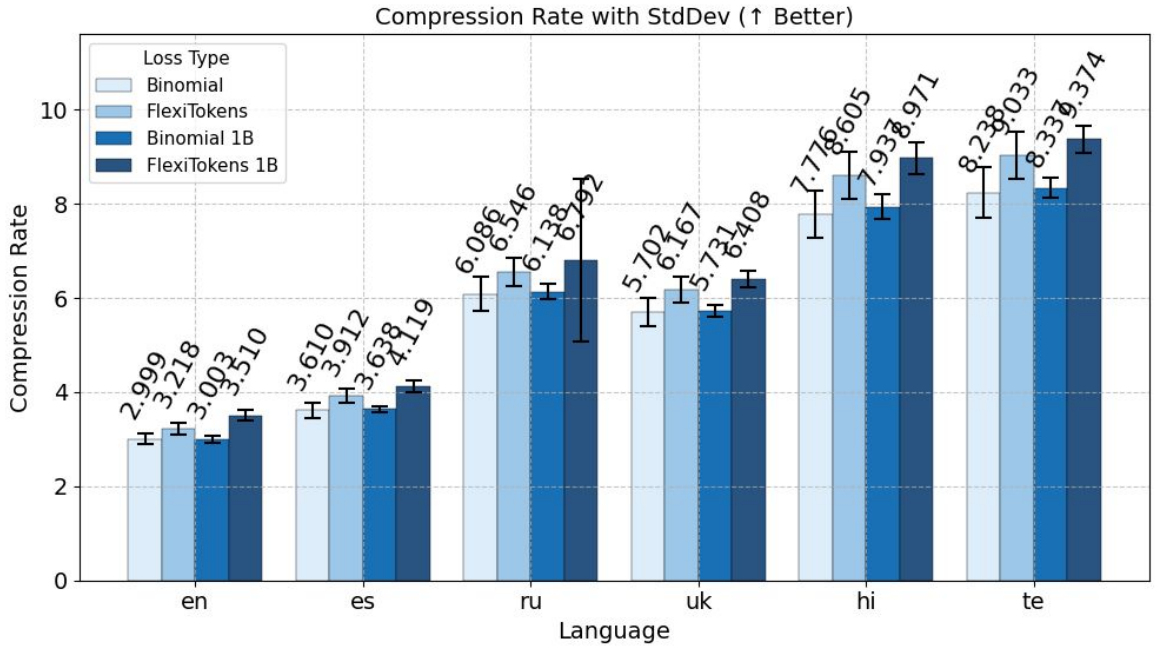




Pretraining Results:

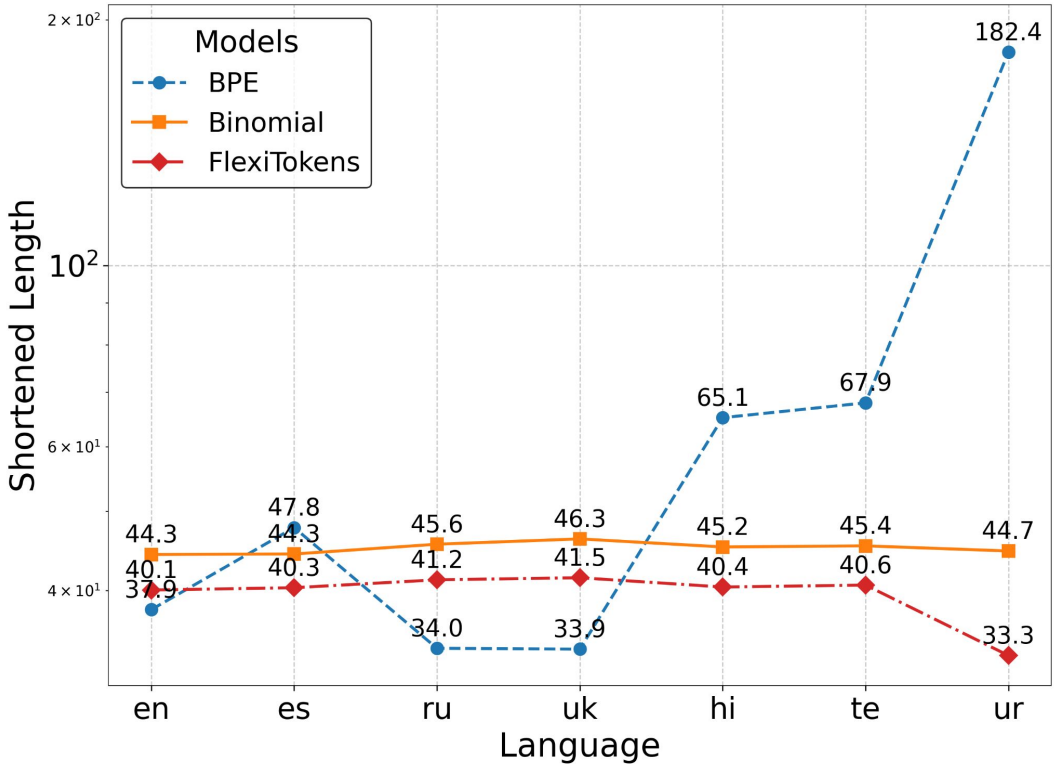
- Achieve comparable BPC with even higher compression rate
  - Implication: higher compute efficiency & runtime
- Higher variance: higher flexibility in how input sequences are fragmented





Pretraining Results:

- FlexiTokens consistently produces the least number of tokens while maintaining balance across languages, even for the unseen language Urdu.
- BPE over-fragments seen (Hindi, Telugu) as well as unseen languages (Urdu).



Average number of tokens per sample obtained in the FLORES dataset

Downstream Results:

Table 3: WikiANN (NER), XNLI and SIB-200 F1 Score and Accuracy and for 3× Compression Rate. FLEXITOKENS outperforms all baselines on XNLI and NER respectively. Notably, it achieves approximately a 3 point gain on XNLI for Urdu—an unseen language script—compared to BPE.

	NER F1 Score						
Model	**en**	**es**	**ru**	**uk**	**hi**	**te**	**Avg**
BPE	52.30	67.7 0	64.94	74.99	60.23	48.18	61.39
BINOMIAL	63.80	75.06	67.59	**78.06**	61.21	48.31	65.67
FLEXITOKENS $\lambda 1$	63.07	76.12	**68.30**	77.94	**62.26**	**51.74**	**66.57**
FLEXITOKENS $\lambda 2$	**63.96**	**76.23**	67.55	77.99	62.24	48.13	66.02
FLEXITOKENS $\lambda 3$	63.73	75.45	68.25	78.01	61.97	50.88	66.38
FLEXITOKENS $\lambda 3$ **1B**	64.61	77.60	69.69	79.53	63.61	52.77	67.97

NER: FlexiTokens outperforms other baselines

Downstream Results:

Table 3: WikiANN (NER), XNLI and SIB-200 F1 Score and Accuracy and for 3× Compression Rate. FLEXITOKENS outperforms all baselines on XNLI and NER respectively. Notably, it achieves approximately a 3 point gain on XNLI for Urdu—an unseen language script—compared to BPE.

NER F1 Score							
Model	**en**	**es**	**ru**	**uk**	**hi**	**te**	**Avg**
BPE	52.30	67.7 0	64.94	74.99	60.23	48.18	61.39
BINOMIAL	63.80	75.06	67.59	**78.06**	61.21	48.31	65.67
FLEXITOKENS $\lambda 1$	63.07	76.12	**68.30**	77.94	**62.26**	**51.74**	**66.57**
FLEXITOKENS $\lambda 2$	**63.96**	**76.23**	67.55	77.99	62.24	48.13	66.02
FLEXITOKENS $\lambda 3$	63.73	75.45	68.25	78.01	61.97	50.88	66.38
FLEXITOKENS $\lambda 3$ **1B**	64.61	77.60	69.69	79.53	63.61	52.77	67.97

Compression Rate ± Std							
Binomial 3x	3.05 ± 0.47	3.88 ± 0.76	6.37 ± 1.67	5.75 ± 1.11	8.74 ± 3.27	8.56 ± 2.29	6.06 ± 1.86
FLEXITOKENS $\lambda 1$	3.18 ± 0.43	3.84 ± 0.54	6.31 ± 1.15	5.92 ± 0.90	8.42 ± 1.68	8.64 ± 1.55	6.05 ± 1.14
FLEXITOKENS $\lambda 2$	3.27 ± 0.44	3.93 ± 0.58	6.58 ± 1.38	6.12 ± 1.00	8.52 ± 1.49	9.15 ± 2.21	5.66 ± 1.33
FLEXITOKENS $\lambda 3$	3.42 ± 0.53	4.18 ± 0.66	6.64 ± 1.29	6.30 ± 1.07	8.76 ± 1.77	8.99 ± 2.07	6.38 ± 1.35

NER: FlexiTokens outperforms other baselines

Downstream Results:

Table 3: WikiANN (NER), XNLI and other tasks' F1 Score and Accuracy and for 3× Compression Rate. FLEXITOKENS outperforms all baselines on XNLI and NER respectively. Notably, it achieves approximately a 3-point gain on XNLI for Urdu——an unseen language script——compared to BPE

	XNLI Accuracy						
Model	**en**	**es**	**ru**	**hi**	**te**	**ur** (OOD)	**Avg**
BPE	73.09	69.9	65.95	61.48	**68.00**	54.11	65.42
BINOMIAL	72.87	70.28	65.93	62.26	66.11	54.79	65.37
FLEXITOKENS $\lambda 1$	**73.51**	70.22	66.47	**62.42**	67.11	56.99	66.12
FLEXITOKENS $\lambda 2$	73.21	**70.84**	**66.97**	62.16	66.71	**57.58**	**66.25**
FLEXITOKENS $\lambda 3$	73.35	70.22	66.75	62.36	67.82	57.33	66.31
FLEXITOKENS $\lambda 3$ **1B**	75.17	72.44	68.60	64.41	69.62	57.62	67.98

XNLI: FlexiTokens outperforms other baselines

Downstream Results:

Table 3: WikiANN (NER), XNLI and other tasks' F1 Score and Accuracy and for 3× Compression Rate. FLEXITOKENS outperforms all baselines on XNLI and NER respectively. Notably, it achieves approximately a 3-point gain on XNLI for Urdu——an unseen language script——compared to BPE

SIB-200 Accuracy							
Model	**en**	**es**	**ru**	**uk**	**hi**	**te**	**Avg**
BPE	**80.88**	**81.37**	**81.37**	**76.96**	60.78	**72.55**	**75.65**
BINOMIAL	79.41	74.02	71.08	68.63	64.71	69.61	71.24
FLEXITOKENS $\lambda 1$	78.92	72.55	75.49	69.61	61.27	66.18	70.67
FLEXITOKENS $\lambda 2$	77.94	75.98	74.51	71.57	69.12	66.18	72.55
FLEXITOKENS $\lambda 3$	**80.88**	77.45	73.04	72.55	**71.08**	**71.08**	74.35
FLEXITOKENS $\lambda 3$ **1B**	85.78	83.82	86.27	84.31	77.94	81.86	83.33

SIB: FlexiTokens outperforms Binomial baselines

Downstream Results:

Table 3: WikiANN (NER), XNLI and other tasks' F1 Score and Accuracy and for 3× Compression Rate. FLEXITOKENS outperforms all baselines on XNLI and NER respectively. Notably, it achieves approximately a 3-point gain on XNLI for Urdu——an unseen language script——compared to BPE

	Med. Abs./Irony/CS/CS/ILI - Accuracy					
Model	**Med. Abs. en**	**Irony en**	**CS en-es**	**CS en-hi**	**ILI hi**	**Avg**
BPE	57.68	67.86	**92.48**	**87.36**	89.06	78.89
BINOMIAL	62.81	67.60	91.62	84.98	89.47	79.30
FLEXITOKENS $\lambda 1$	62.92	68.37	-	-	89.58	73.62
FLEXITOKENS $\lambda 2$	62.74	68.75	91.37	86.53	**90.33**	79.94
FLEXITOKENS $\lambda 3$	**63.19**	**69.26**	92.11	86.41	89.55	**80.10**

Others: FlexiTokens is competitive with other baselines

Downstream Results:

Table 2: COMET scores on OPUS-100 machine translation task. FLEXITOKENS outperforms across all languages.

Model	en-es	en-ru	en-hi
BPE	59.46	52.53	50.94
BINOMIAL	63.05	57.33	54.35
FLEXITOKENS $\lambda 3$	**64.08**	**57.76**	**54.73**

Translation: FlexiTokens achieves better translation performance

Ablation Results

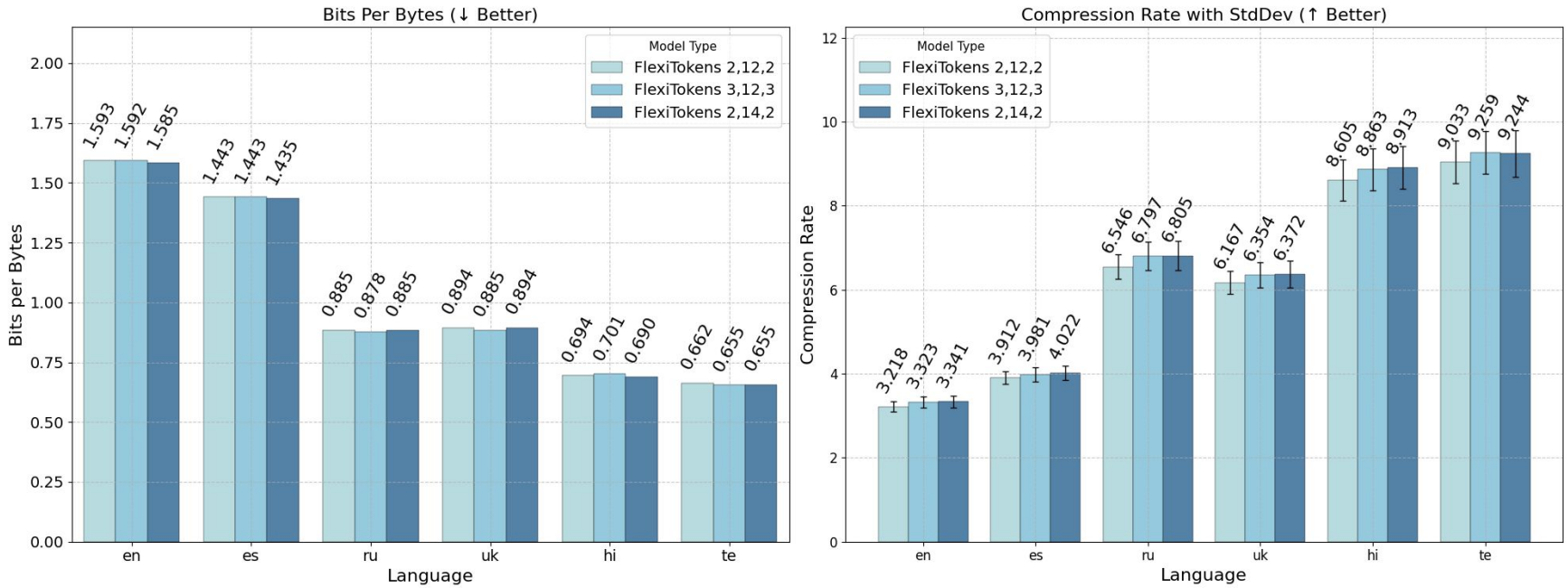
- **Compression has its limits**

Table 4: Ablation for α: Average Accuracy and Compression Results Across Multiple Languages

Model	SIB-200	WikiANN	Multi. Senti.	XNLI	ILI	Med. Abs.	Avg
Accuracy							
FLEXITOKENS 10x	53.76	64.35	**72.99**	65.23	89.07	62.95	68.06
FLEXITOKENS 5x	71.16	64.92	72.54	65.48	89.28	**63.47**	71.14
FLEXITOKENS 3x	**72.55**	**66.02**	72.74	**66.25**	**90.33**	62.74	**71.77**
Compression Rate ± Std							
FLEXITOKENS 10x	28.89 ± 11.06	28.01 ± 14.14	27.41 ± 12.12	29.06 ± 8.55	38.80 ± 38.80	13.22 ± 2.15	27.56 ± 14.47
FLEXITOKENS 5x	10.72 ± 1.54	11.17 ± 3.69	11.25 ± 2.86	12.15 ± 1.76	14.82 ± 14.82	5.63 ± 0.33	10.96 ± 4.17
FLEXITOKENS 3x	6.19 ± 0.53	6.26 ± 1.33	6.17 ± 1.03	6.83 ± 0.60	8.35 ± 8.35	3.21 ± 0.15	6.17 ± 2.00

Scaling Results

- **More layers, better BPC**
  - **LM submodule > others**



Analysis Results

- **BPE breaks down in unseen languages.**
- **With FlexiTokens, we more semantically meaningful tokenization patterns**

Tokenizer	Sentence and Segmentation	#Tokens
ur	39-year-old SpongeBob was diagnosed with hypertrophic cardiomy-opathy in Mumbai.	–
BPE	39|Ø|³|ØŞÙĦ|Û|ģ| |ØŞ|Ø|³|Ù¾|ÙÎ|Ø|¬| Ø|¨|ØŞ|Ø|¨| |Ú©|ÙĪ| Ùħ|Ùħ|Ø|¨|Ø|¦|Û|Į| Ùħ|Û|Į|Úº| |Û|ģ|ØŞ|Ø|¦|Ù¾|Ø±|Ù¹|Ø±|ØŞ|Ù|ģ|Ú©| |Ú©|ØŞ|Ø±|ÚĪ|Û|Į|ÙĪ|Ùħ|Û|Į|ÙĪ|Ù¾|Û|Į|Ø|ª|Ú¾|Û|Į| |Ú©|Û|Į| Ø|ª|Ø|´|Ø|®|Û|Į|Ø|µ| |Û|ģ|ÙĪ|Ø|¦|Û|Į|Û|Ķ	107
BINOMIAL 3×	3|9 |سالہ |اسپنج |باب |کو |ممبئ |میں |ھائپر|ٹرافک |کارڈیو|میو|پیتھی |کی |تشخ|ص |ھوائی|-	21
FLEXITOKENS 3×	9 3|سالہ |اسپنج |باب |کو |ممبئی |میں |ھائپرٹرافک |کارڈیو|میو|پ|تھ| |کی |تشخیص |ھوئی|-	17

3:9 :year:old :Sponge :Bob :to :Mumbai :in :**hyper**:tro:phic :**cardio**:**myo**:pathy :of :diag:no:sis :was:done.

Analysis Results:

- **No tradeoffs with English**

en	Influenza and pneumonia were identified as major causes of mortality in children.	–
BPE	In|f|lu|enza| and| p|ne|um|onia| were| identified| as| major| causes| of| mort|ality| in| children|.	20
BINOMIAL 3×	I|n|fl|uenza |an|d |pn|eumon|ia |wer|e |id|ent|ified |as |maj|or| |causes |of |mor|t|ality |in| |chil|dren.	25
FLEXITOKENS 3×	Infl|uenz|a |and| |pneu|m|onia |were |identified |as |m|ajor| |causes |of |m|or|tality |in |childr|en.	20

Analysis Results

- **Emojis crash the BPE tokenizer**

en	Oh no, another surprise bonus at work. Just what I didn't need 😀😂😂🙋🏽‍♂️🤸‍♀️.	–
BPE	Oh| no|,| another| surprise| bonus| at| work|.| Just| what| I| didn|âĢĻt| need| ðŁĺ|Ģ|ðŁĺ|Ĥ|ðŁĺ|Ĥ|ðŁ|Ļ|ĺ|ðŁ|ı|½|âĢ|ı|âĻ|Ĥ|ï¸ı|ðŁ|¤¸|ðŁ|ı|¾|âĢ| ı|âĻ|Ģ|ï¸ı|ł|ħ|§|ł|ħ|ķ|ł|Ħ|IJ|ł|ħ|ij|ł|ħ|¢|ł|ħ|ķ|ł|Ħ|IJ|ł|ħ|ĵ|ł|ħ|Ł|ł|ħ|Ł|ł|ħ|Ľ|ł|ħ|ķ|ł|ħ|Ķ|.	82
BINOMIAL 3×	Oh no, |an|ot|her| |sur|pr|ise |bon|us |at| |w|ork. |Just| |w|hat |I didn'|t| |need | 😀|||||😂😂| 🙋🏽‍♂️||||🤸‍♀️|.	33
FLEXITOKENS 3×	Oh |no, |anot|her |sur|pr|ise |bonus |at| |wor|k.| |J|ust |what| |I |didn|t| |need | 😀😂😂| 🙋🏽‍♂️| 🤸‍♀️|.	25

Takeaways

- **Current tokenization methods are not flexible as we change domains.**
- **Allow the model find its best tokenization pattern, you get better results.**
- **Learnable tokenization like FlexiTokens makes tokenization fair across languages**
- **Increased inference speed with fewer tokens.**
- **Future directions involve more scaling to larger sizes**

Thank you

Collaboration and helpful feedback?
Please reach out: Owodunni.1@osu.edu

MiSc: Why Less Tokens is better

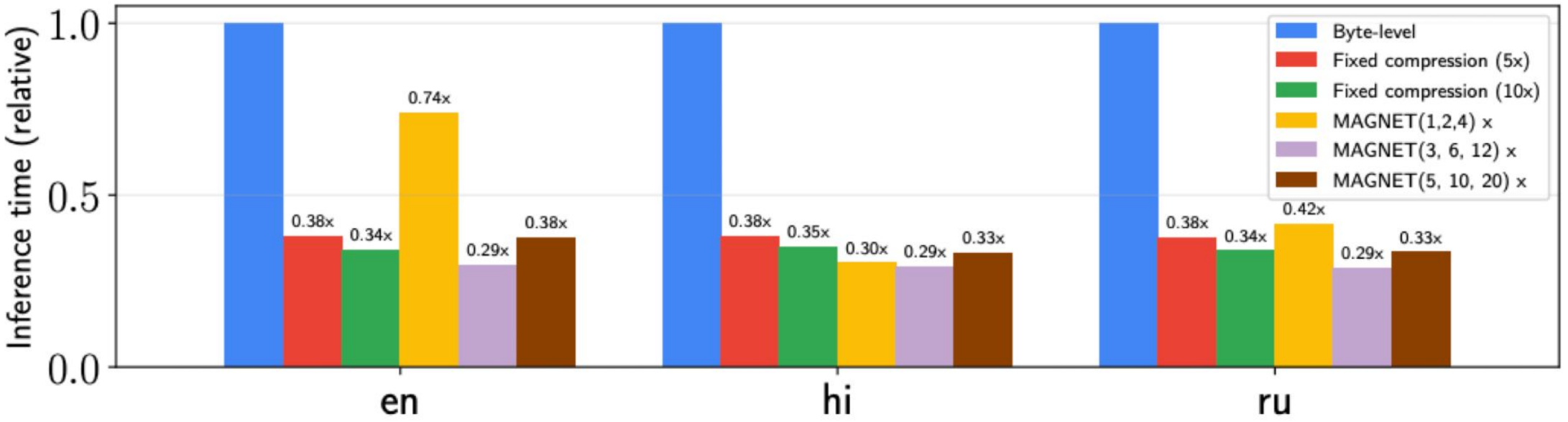


Figure 4: Inference time per language in XQUAD, relative to the byte-level model. MAGNET's inference time is shorter than the byte-level model and comparable to DTP for most of the languages.