Where is the signal in tokenization space?

Renato Geh, Honghua Zhang, Kareem Ahmed,
Benjie Wang, Guy Van den Broeck

University of California, Los Angeles

Tokenization

Most language models represent distributions over sequences of *tokens* (subwords), not strings.

$$\begin{aligned} \textbf{string} \quad & \mathbf{x} = (x_1, x_2, \ldots, x_n) \\ \textbf{tokenization} \quad & \mathbf{v} = (v_1, \ldots, v_m) \end{aligned}$$

For example:

$$\begin{aligned} \textbf{string} \quad & \mathbf{x} = \texttt{Caterpillar} \\ \textbf{tokenization} \quad & \mathbf{v} = [\texttt{C},\texttt{ater},\texttt{p},\texttt{ill},\texttt{ar}] \equiv [\texttt{315},\texttt{1008},\texttt{29886},\texttt{453},\texttt{279}] \end{aligned}$$

Canonical Tokenization

How do we tokenize? There is usually a unique *canonical* tokenization:

$$\begin{aligned} \textbf{string} \quad \mathbf{x} &= \texttt{Caterpillar} \\ \textbf{canonical} \quad \mathbf{v} &= [\texttt{C},\texttt{ater},\texttt{p},\texttt{ill},\texttt{ar}] \end{aligned}$$

(Llama 2)

Common assumption:

$$p(\mathbf{x}) = p(\mathbf{v})$$

✘

A string can be tokenized in an exponential number of ways (784 here!)

$$\begin{gathered} [\texttt{C},\texttt{ater},\texttt{pi},\texttt{l},\texttt{lar}], [\texttt{Cat},\texttt{er},\texttt{pi},\texttt{lla},\texttt{r}], [\texttt{Cat},\texttt{er},\texttt{pi},\texttt{l},\texttt{lar}], \\ [\texttt{Ca},\texttt{ter},\texttt{p},\texttt{ill},\texttt{ar}], [\texttt{Ca},\texttt{ter},\texttt{p},\texttt{illa},\texttt{r}], [\texttt{Cat},\texttt{er},\texttt{pi},\texttt{ll},\texttt{ar}], \\ \ldots \\ [\texttt{Ca},\texttt{t},\texttt{e},\texttt{r},\texttt{p},\texttt{i},\texttt{l},\texttt{l},\texttt{a},\texttt{r}], [\texttt{C},\texttt{a},\texttt{t},\texttt{e},\texttt{r},\texttt{p},\texttt{i},\texttt{l},\texttt{l},\texttt{a},\texttt{r}] \end{gathered}$$

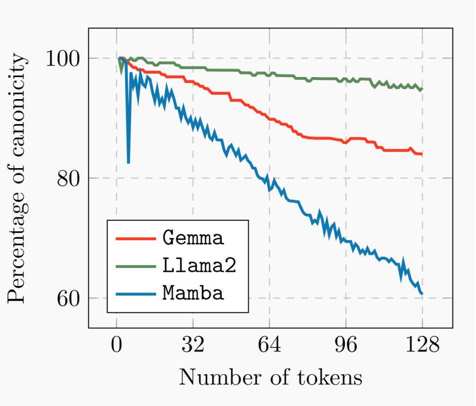
(Llama 2)

Tokenization

Less likely for non-English (code, unicode characters, etc)

Why does this tokenization problem matter?

How canonical are unconditional samples?



string $\mathbf{x} = \texttt{Hypnopaturist}$

canonical $\mathbf{v} = \texttt{[Hyp,nop,atu,rist]}$

most likely $\mathbf{v} = \texttt{[Hyp,no,patu,rist]}$

canonical prob $p(\mathbf{v}|\mathbf{x}) \approx 0.0004$

most likely prob $p(\mathbf{v}|\mathbf{x}) \approx 0.9948$

(Gemma 2B)

We're ignoring an exponential number of tokenizations!

Tokenization is a Neurosymbolic Problem!

↪ Tokens are symbols.
↪ A tokenization of a text is a constraint over these symbols.

$$p(\mathbf{v}, \mathbf{x}) = \begin{cases} p_{\mathrm{LLM}}(\mathbf{v}) & \text{if } \mathbf{v} \models \mathbf{x}; \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbf{v} = (v_1, v_2, \ldots, v_n) \models \mathbf{x} \Leftrightarrow v_1 \circ v_2 \circ \cdots \circ v_n = \mathbf{x}$$

concatenation

Example:

$$p(\mathbf{v} = [\text{一　ラ}, \text{ク}] \,|\, \mathbf{x} = \text{一ラク}) = 0.586 \qquad p(\mathbf{v} = [\text{一}, \text{ラク}] \,|\, \mathbf{x} = \text{一ラク}) = 0.402$$

$$p(\mathbf{v} = [\text{一}, \text{ラ}, \text{ク}] \,|\, \mathbf{x} = \text{一ラク}) = 0.012 \qquad p(\mathbf{v} = [\texttt{Tok}, \texttt{ens}] \,|\, \mathbf{x} = \text{一ラク}) = 0$$

From Gemma 2B

Reasoning in Tokenization Space

Instead of the *canonical* tokenization, we might want to compute:

1. The most likely tokenization ✘

$$\arg\max_{\mathbf{v} \models \mathbf{x}} p(\mathbf{v}, \mathbf{x})$$

Theorem. *The most likely tokenization problem is NP-hard.*

For autoregressive models, e.g. transformers and state space models

2. The true probability of a text ✘

$$p(\mathbf{x}) = \sum_{\mathbf{v} \models \mathbf{x}} p(\mathbf{v}, \mathbf{x})$$

Theorem. *The marginal string probability problem is #P-hard.*

(Approximate) Reasoning in Tokenization Space

1. The most likely tokenization

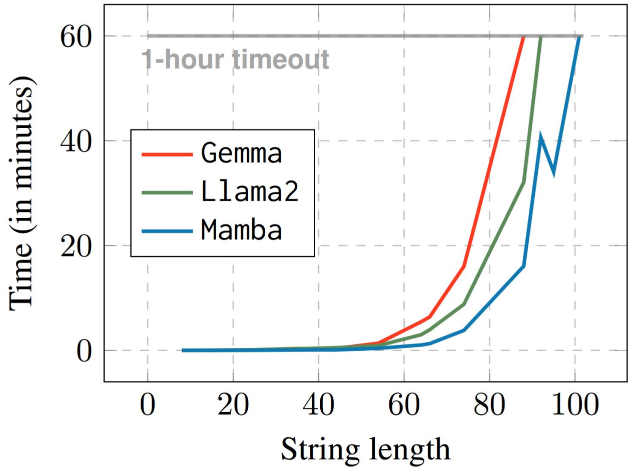
$$\arg\max_{\mathbf{v} \models \mathbf{x}} p(\mathbf{v}, \mathbf{x})$$

Branch-and-bound

- ↪ Lower bound: canonical likelihood
- ↪ Anytime: candidate at least as good as canonical

What did we learn?

- ↪ Runtime exponential on string length!
- ↪ Canonical best candidate for almost all cases…



…not always!

$p(\mathbf{v} = [\texttt{␣tongue,less}] \mid \mathbf{x} = \texttt{␣tongueless}) = 0.518$ → most likely tokenization

$p(\mathbf{v} = [\texttt{␣t,ong,uel,ess}] \mid \mathbf{x} = \texttt{␣tongueless}) = 0.004$

$p(\mathbf{v} = [\texttt{␣tong,uel,ess}] \mid \mathbf{x} = \texttt{␣tongueless}) = 0.474$

canonical tokenization

$p(\mathbf{v} = [\texttt{␣,HEADER,\_,DELIM,ITER}] \mid \mathbf{x} = \texttt{␣HEADER\_DELIMITER}) = 0.412$

$p(\mathbf{v} = [\texttt{␣HEAD,ER,\_,DELIM,ITER}] \mid \mathbf{x} = \texttt{␣HEADER\_DELIMITER}) = 0.330$

$p(\mathbf{v} = [\texttt{␣HEADER,\_,DELIM,ITER}] \mid \mathbf{x} = \texttt{␣HEADER\_DELIMITER}) = 0.010$

canonical tokenization

(Gemma 2B)

(Approximate) Reasoning in Tokenization Space

2. The true probability of a text

$$p(\mathbf{x}) = \sum_{\mathbf{v} \models \mathbf{x}} p(\mathbf{v}, \mathbf{x})$$

Sequential importance sampling

$$p(\mathbf{x}) = \mathbb{E}_{\mathbf{v} \sim q(\mathbf{v}|\mathbf{x})}\left[\frac{p(\mathbf{v}, \mathbf{x})}{q(\mathbf{v}|\mathbf{x})}\right]$$
$$\approx \frac{1}{N}\sum_{i=1}^{N} \frac{p\left(\mathbf{v}^{(i)}, \mathbf{x}\right)}{q\left(\mathbf{v}^{(i)}|\mathbf{x}\right)}$$

Unbiased estimator converging to the true probability of text as #samples grows

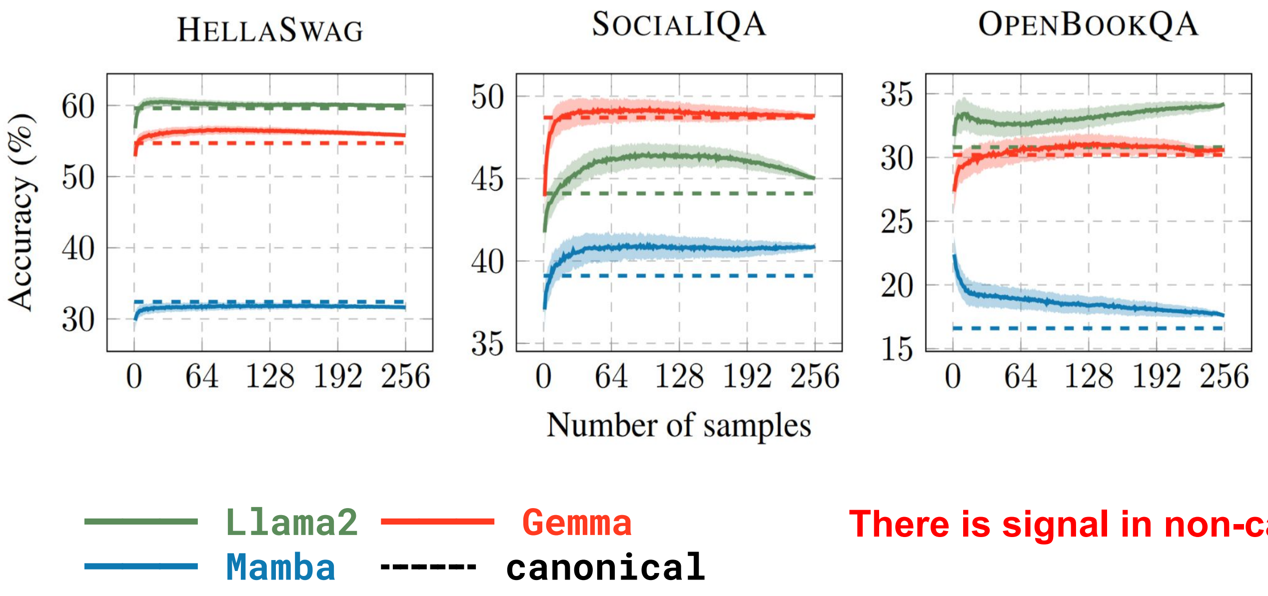
proposal distribution

$$q(v_j|\mathbf{v}_{1:j-1} = [\texttt{Tok,eni}], \mathbf{x} = \texttt{Tokenization}) = \begin{cases} 0.50 & \text{, if } v_j = \texttt{zation}; \\ 0.30 & \text{, if } v_j = \texttt{zat}; \\ 0.15 & \text{, if } v_j = \texttt{za}; \\ 0.05 & \text{, if } v_j = \texttt{z}; \\ 0.00 & \text{, if } v_j = \texttt{a}; \\ \vdots & \\ 0.00 & \text{, if } v_j = \texttt{zzz}; \end{cases}$$

zero-out next tokens inconsistent with constraint

Where is the signal in tokenization space?

$$\arg\max_{\text{answer}} \sum_{\mathbf{v} \models \text{answer}} p(\mathbf{v}, \text{answer} | \mathbf{v}_{\text{question}})$$



California experiences heavy earthquake activity due to

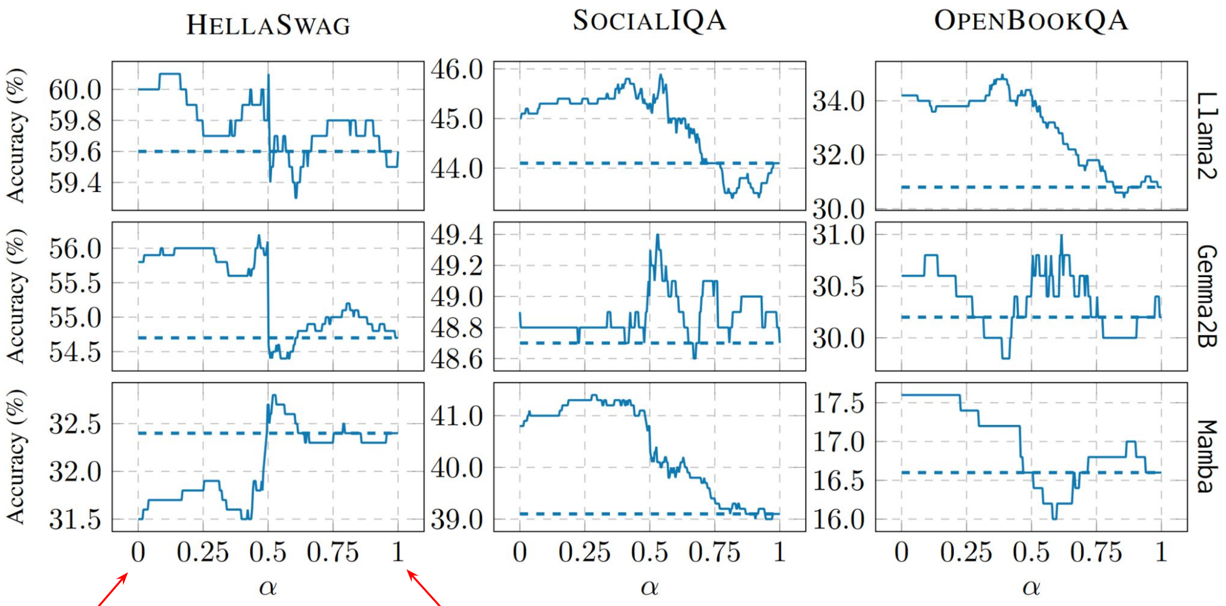
- (a) erosion
- (b) techtonics
- (c) volcanic activity
- (d) fire

There is signal in non-canonical tokenizations!

Mixtures of tokenizations can boost LLM accuracy!

Can we quantify how much signal is in non-canonical tokenizations?

$$\arg\max_{\text{answer}} \alpha \cdot \underbrace{p(\mathbf{v}_{\text{answer}} | \mathbf{v}_{\text{question}})}_{\text{canonical}} + (1-\alpha) \cdot \underbrace{p(\text{noncanonical} | \mathbf{v}_{\text{question}})}_{\text{non-canonicals}}$$



only non-canonicals

only canonical

Tune for α

	Mixture	Canonical	
	Accuracy (%)		
Llama2	**59.7**	59.6	HellaSwag
Gemma	**55.8**	54.7	
Mamba	31.6	**32.4**	
Llama2	**44.8**	44.1	SocialIQA
Gemma	**48.8**	48.7	
Mamba	**39.8**	39.1	
Llama2	**34.0**	30.8	OpenBookQA
Gemma	**30.6**	30.2	
Mamba	**17.6**	16.6	

Consistent improvement!

Main Takeaways

Probabilistic reasoning is hard

✗ Computing the most likely tokenization (exactly) is **hard**

✗ Computing the true text probability (exactly) is **hard**

Non-canonical tokenizations appear in the wild

✔ LLMs sample non-canonical tokenizations

✔ Non-canonical tokenizations can be more likely

Non-canonical tokenizations matter

✔ Mixtures of canonical and non-canonical boost performance

✔ More inference time compute, better performance

Where is the signal in tokenization space?

Renato Geh, Honghua Zhang, Kareem Ahmed,
Benjie Wang, Guy Van den Broeck

University of California, Los Angeles

UCLA

STAR AI
RESEARCH LAB
UCLA