Where is the signal in [token,iza,tion] space?

Renato Lui Geh, Honghua Zhang, Kareem Ahmed, Benjie Wang, Guy Van den Broeck
University of California, Los Angeles
{renatolg,hzhang19,ahmedk,benjiewang,guyvdb}@cs.ucla.edu

(TL;DR) Does the signal come from only one source, i.e. the so-called **canonical** (default) **tokenization**?

The answer is **no**!

By looking at **non-canonical tokenizations**, we get **consistent improvement** in downstream performance! 👇

(1)

Given a string, what's $p(x)$ under the LLM?

string x = `Caterpillar`
tokenization v = `[C,ater,p,ill,ar]`

Common assumption:

$$p(x) = p(v) \quad ✗$$

What about *other* tokenizations?

`[C,ater,pi,l,lar]`, `[Cat,er,pi,lla,r]`, …, `[C,at,e,r,p,i,l,l,a,r]`

are all valid tokenizations!

$$p(x) = \sum_v p(v, x) \quad ✓$$

We should not neglect other tokenizations…

(2)

After all, some non-canonical tokenizations can have **non-negligible mass**!

string x = `Hypnopaturist`
canonical v = `[Hyp,nop,atu,rist]`
most likely v = `[Hyp,no,patu,rist]`

canonical prob $p(v|x) \approx 0.0004$
most likely prob $p(v|x) \approx 0.9948$

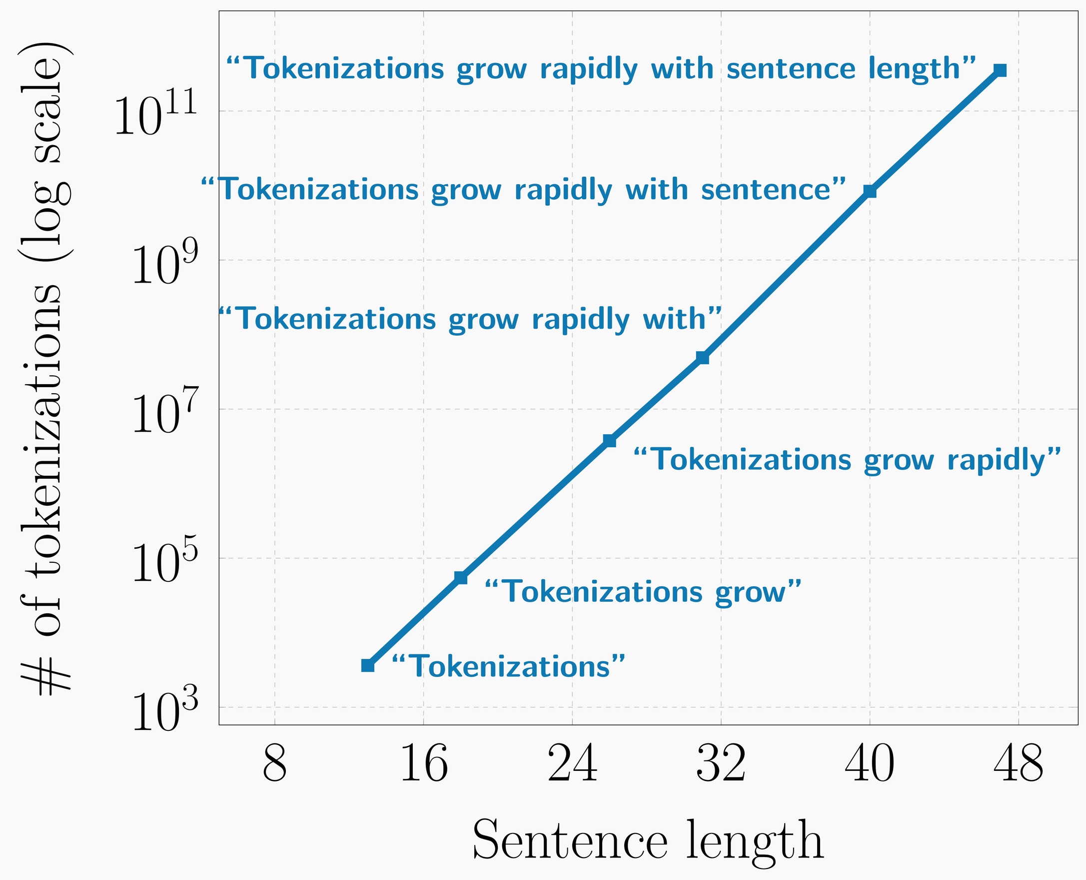
Here, `Gemma`'s **canonical tokenization** v of x is *much less likely* compared to the **most likely** (non-canonical) **tokenization** v.

Read our paper!

`https://arxiv.org/abs/2408.08541`

(3)

…but the space of tokenizations is exponential!



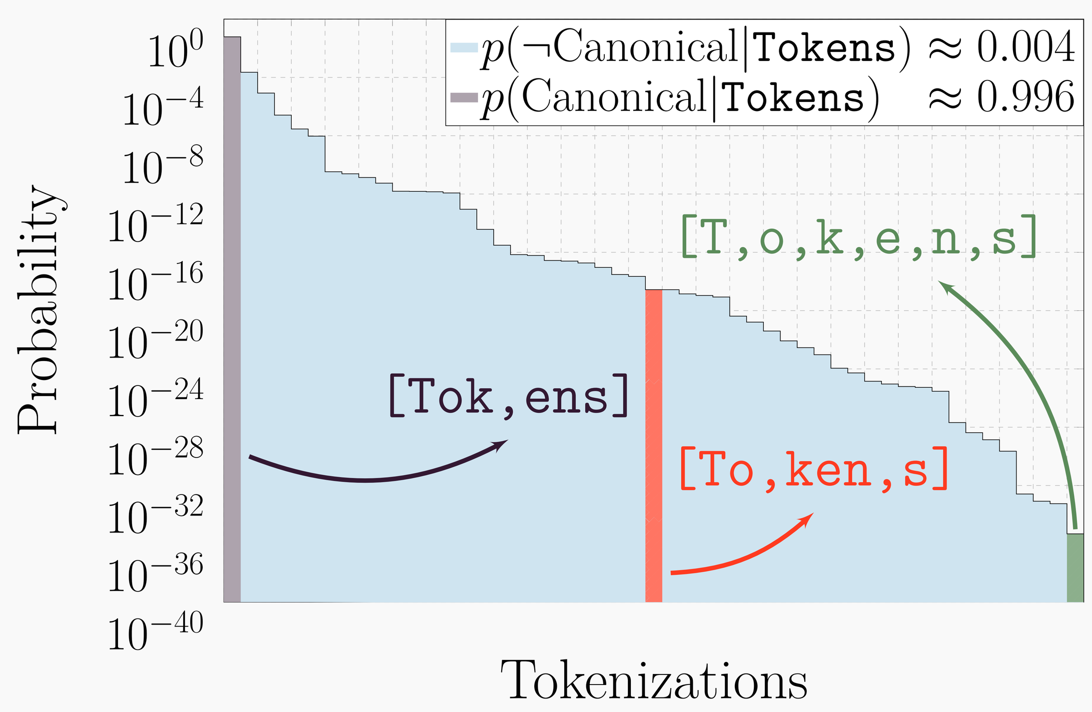
So can we **reason probabilistically** about this tokenization space?

(4)

E.g., exactly compute the **most likely tokenization** for autoregressive models?

Answer: **No!** 😢

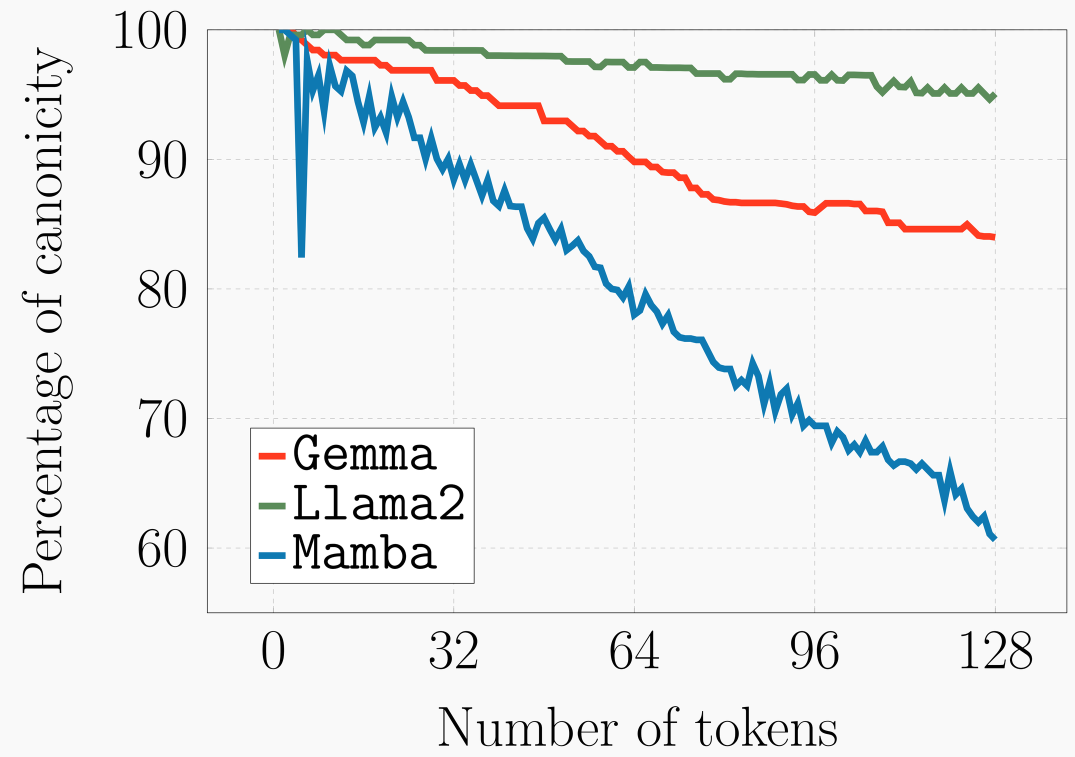
Theorem. *The most likely tokenization problem is NP-hard.*



Yet experimentally, **canonical is often *much more* likely** in English.

(5)

Despite this, **sampling unconditionally** from the LLM reveals that…



…they are generating non-canonically for long texts!

This is especially true for generated non-English (e.g. code, unicode characters, etc.)

(6)

These generated **non-canonical tokenizations** can be more likely than **canonical**!

x = ーラク
$p(v =$ `[ー ラ,ク]` $|x) =$ **0.586**
$p(v =$ `[ー,ラ,ク]` $|x) = 0.012$
$p(v =$ `[ー,ラク]` $|x) = 0.402$
x = `␣tongueless`
$p(v =$ `[␣tongue,less]` $|x) =$ **0.518**
$p(v =$ `[␣t,ong,uel,ess]` $|x) = 0.004$
$p(v =$ `[␣tong,uel,ess]` $|x) = 0.474$
x = `␣HEADER_DELIMITER`
$p(v =$ `[␣,HEADER,_,DELIM,ITER]` $|x) =$ **0.412**
$p(v =$ `[␣HEAD,ER,_,DELIM,ITER]` $|x) = 0.330$
$p(v =$ `[␣HEADER,_,DELIM,ITER]` $|x) = 0.010$

Meaning there is possibly signal in non-canonical tokenizations!

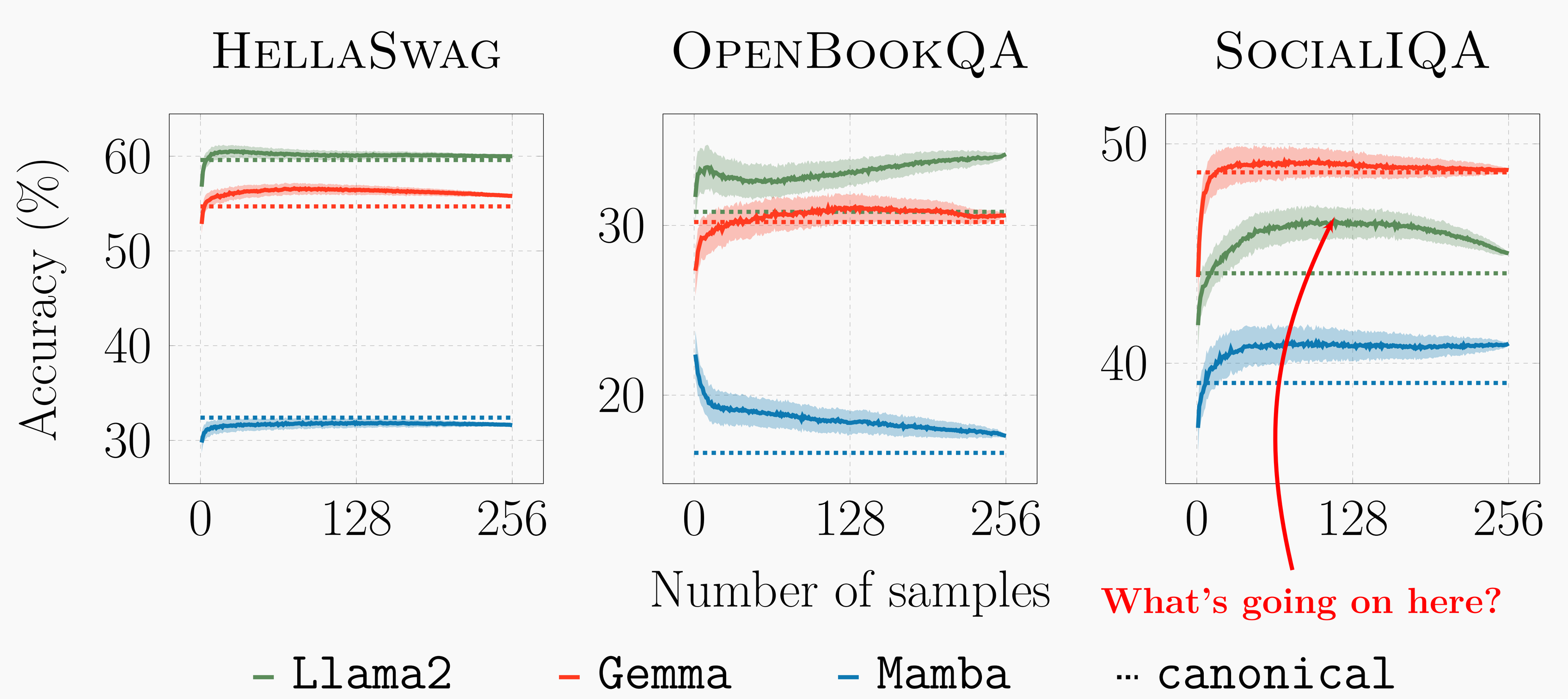
(7)

So can we aggregate over all tokenizations and exactly compute the marginal $p(x) = \sum_v p(v)$?

Answer: **No!** 😢

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

But we *can* approximate through **sequential importance sampling!**

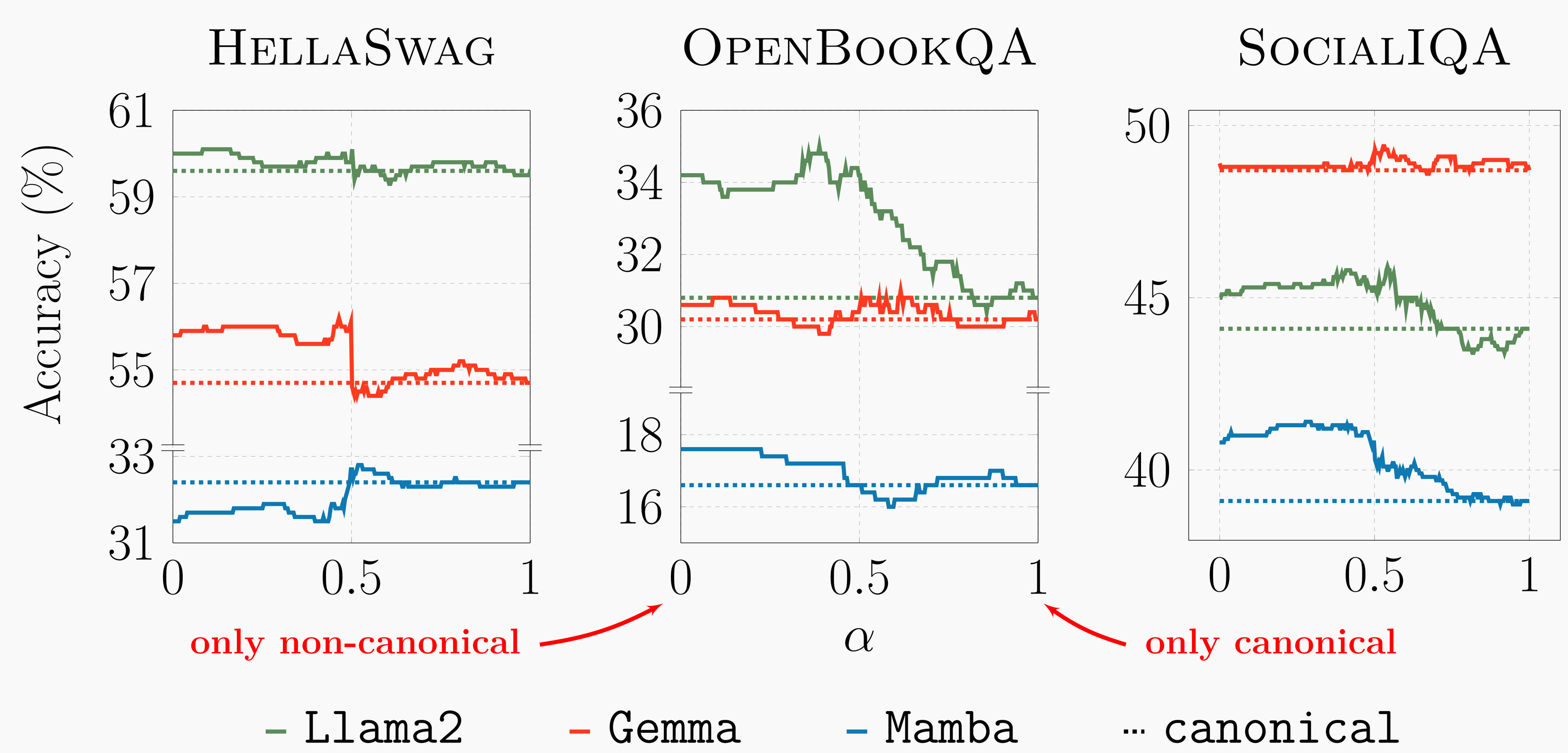


That's evidence of signal in non-canonical tokenizations!

(8)

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

$$\arg\max_{x} \ \alpha \cdot \underbrace{p(v, x|v_q)}_{\text{canonical}} + (1-\alpha) \cdot \underbrace{p(\neg v, x|v_q)}_{\text{non-canonical}}$$



There is **significant signal** in non-canonical tokenizations!