Canonical Autoregressive Generation

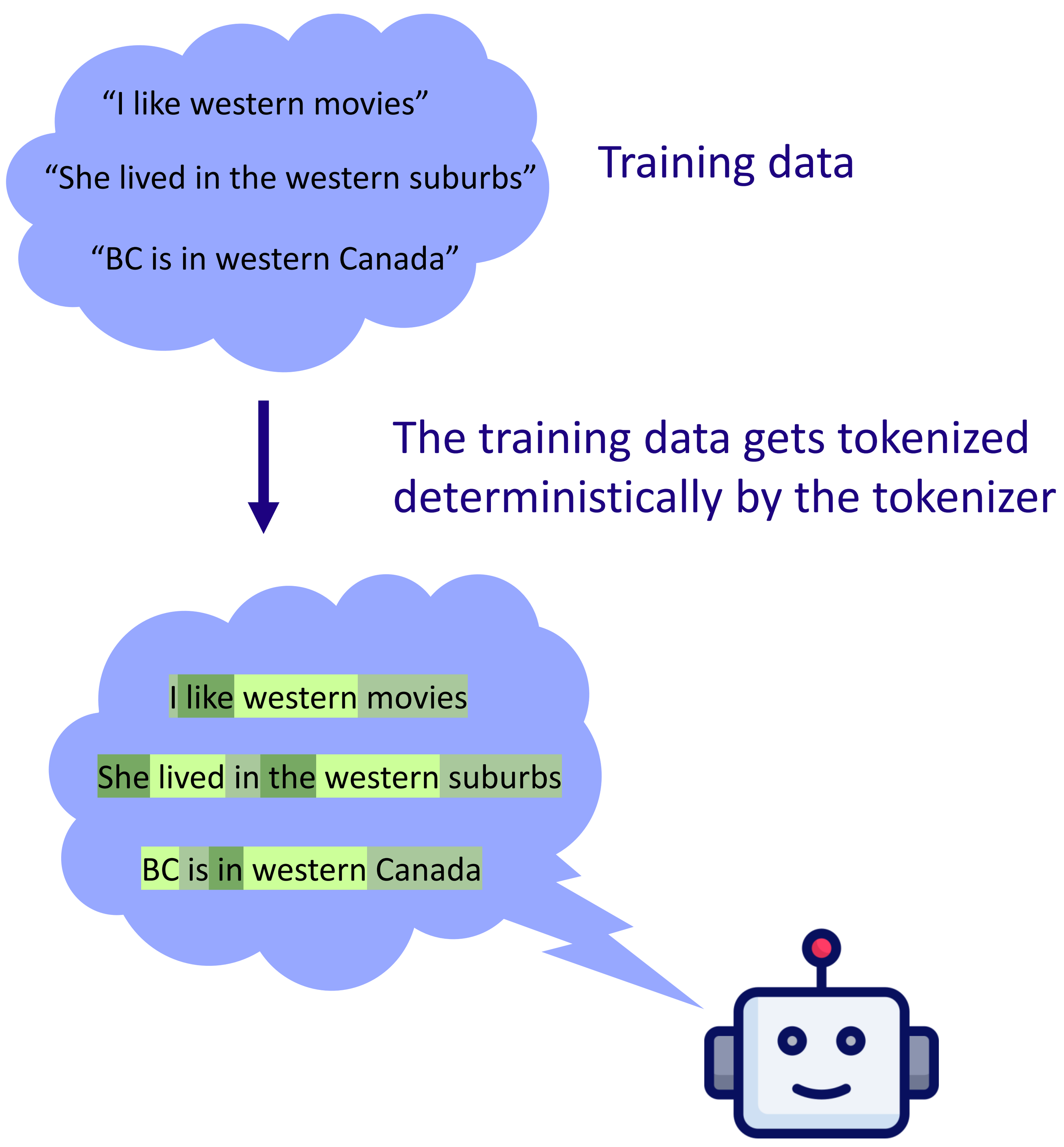
Ivi Chatzi, Nina Corvelo Benz, Stratis Tsirtsis, Manuel Gomez-Rodriguez

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

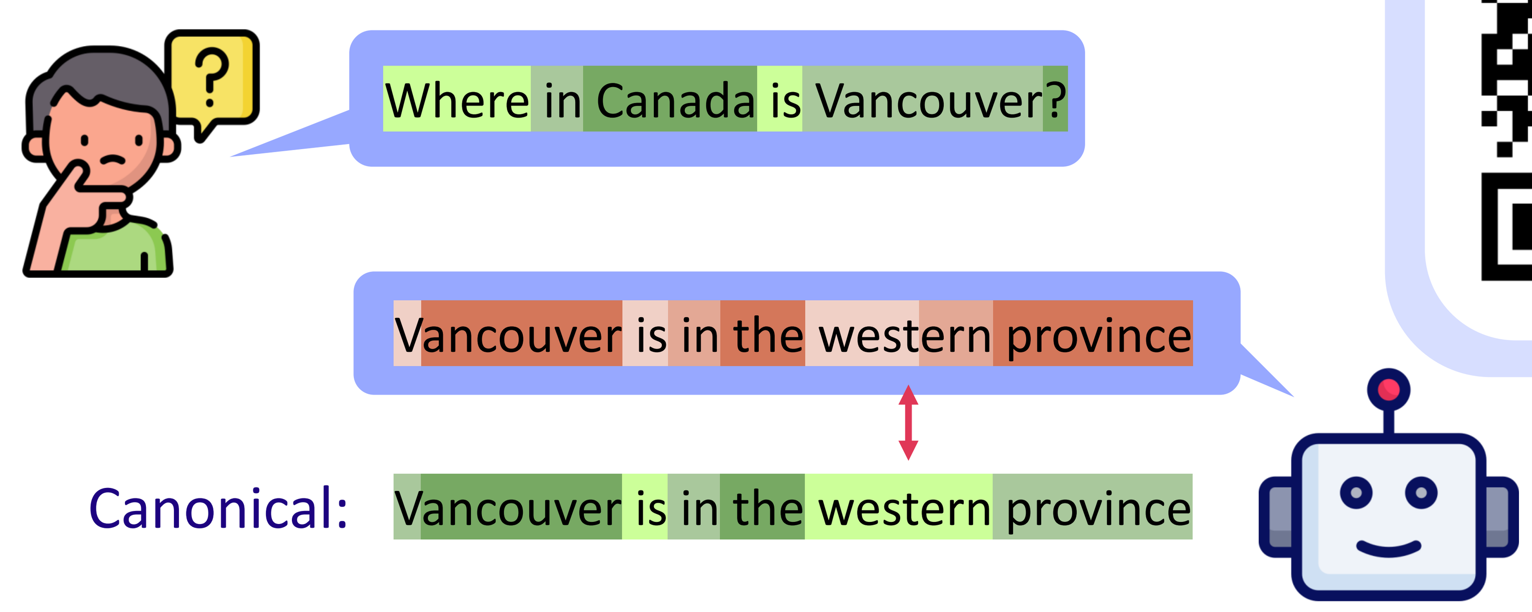
ETH zürich

Paper

LLMs are trained on canonical token sequences



LLMs can generate non-canonical token sequences



Problems with non-canonical token sequences

- They can bypass safety filters leading to harmful responses
  "Adversarial Tokenization" *Geh et al., 2025*
- They allow token misreporting by LLM providers
  "Is Your LLM Overcharging You? Tokenization, Transparency, and Incentives" *Artola Velasco et al., 2025*
- String perplexity is computationally hard
  "Where is the signal in tokenization space?" *Geh et al., EMNLP 2024*
  "You should evaluate your language model on marginal likelihood over tokenisations" *Cao & Rimell, EMNLP 2021*
  "Should you marginalize over possible tokenizations?" *Chirkova et al., ACL 2023*

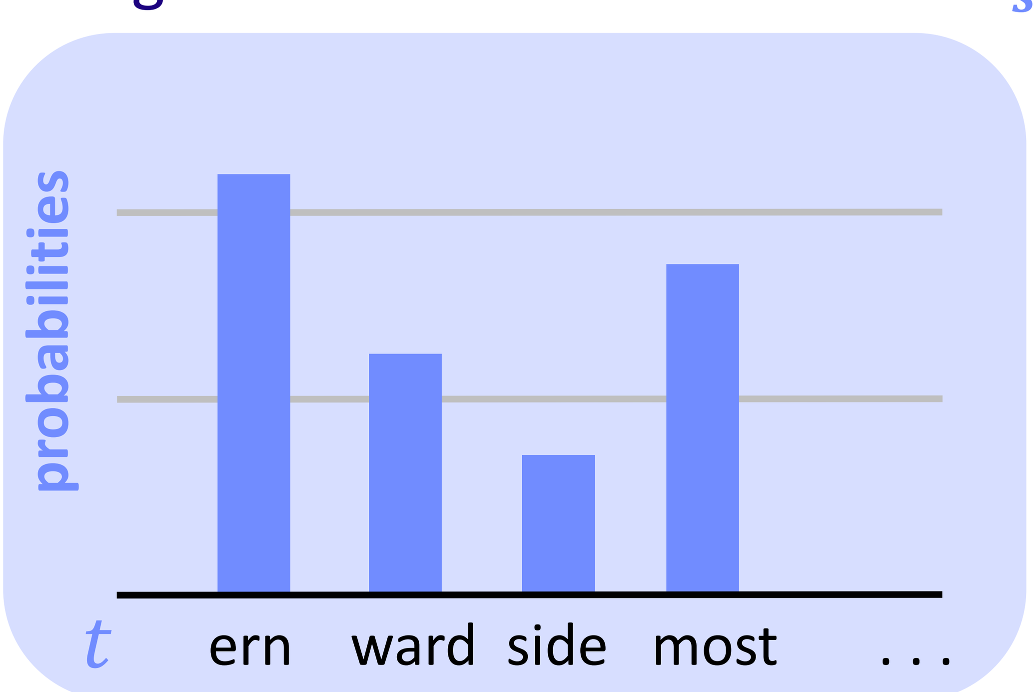
How to ensure LLMs can only generate canonical token sequences?

Theorem
If token sequence s is **non-canonical**, then for any token t the sequence $s \mid t$ is also **non-canonical**

→ At every step of generation the (partial) token sequence generated so far must be **canonical**

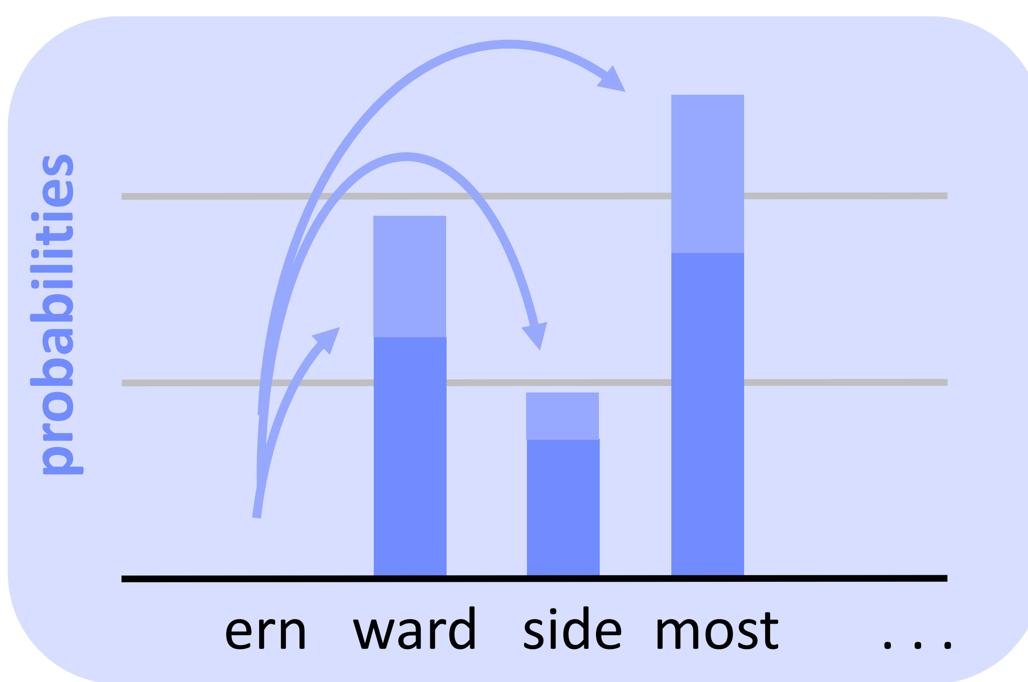
We propose Canonical Sampling

Original next-token distribution d_s



$s|t$ is canonical ✗ ✓ ✓ ✓

Canonicalized next-token distribution $\tilde{d}_s(t) = \begin{cases} \frac{d_s(t)}{Z}, & s \mid t \text{ is canonical} \\ 0, & s \mid t \text{ is non-canonical} \end{cases}$



Theorem
Sampling from $\tilde{d}_s$ leads to output token sequences closer to the true distribution of token sequences (as seen during training)

Efficient Canonical Sampling

Computing the normalization constant $Z = \sum_{t \in V:\ s|t \text{ is canonical}} d_s(t)$ requires checking if $s|t$ is canonical for all tokens t

But we can efficiently sample from $\tilde{d}_s$ using the **Gumbel-Max trick**:

$$\tilde{d}_s(t) \sim \text{argmax}_{t \in V:\ s|t \text{ is canonical}} \{\log(d_s(t)) + u_t\}$$

ALGORITHM

Sample $u_t \sim \text{Gumbel}(0,1)$ for every token t in the vocabulary

For every token t in decreasing order of $\log(d_s(t)) + u_t$:
  If $s|t$ is canonical then return t

Requires fewer than $\frac{1}{Z}$ canonicity checks on average